Phenotypic innovation through recombination in genome-scale metabolic networks

Hosseini, Sayed-Rzgar ; Martin, Olivier C ; Wagner, Andreas

DOI: https://doi.org/10.1098/rspb.2016.1536

Proceedings of the Royal Society B: Biological Sciences

DOI: 10.1098/rspb.2016.1536

Supplementary Information to

Phenotypic innovation through recombination in genome-scale metabolic networks

Sayed-Rzgar Hosseini[1, 2], Olivier C Martin[3] and Andreas Wagner[1, 2, 4]

[1]*Institute of Evolutionary Biology and Environmental Sciences, University of Zurich, Bldg. Y27, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

[2]*The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland*

[3]*GQE-Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, -91190 Gif-sur-Yvette, France*

[4]*The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Text S1:

Supplementary Methods

(a) Genome-scale metabolic networks and their phenotypic representations

The set of genomically encoded biochemical reactions proceeding inside a given organism constitutes an organism's metabolic genotype [1–3]. This genotype enables an organism to extract energy and produce small biomass building blocks, such as amino acids, from extracellular nutrients. Reconstruction of this genotype from genomic and biochemical information has been successful for multiple organisms [4–7].

Each metabolic network contains a subset of the "reaction universe" of all biochemical reactions that take place in the biosphere (See Text S1b). We have curated a representation of this universe, which comprises 5906 reactions and is based on current metabolic knowledge [8–12]. We represent an organism's metabolic genotype as a binary vector of length 5906. Each entry of this vector corresponds to a given reaction in the reaction universe, and is equal to one if the corresponding reaction is present in the metabolic network, and zero otherwise. Thus, each genotype can be thought of as a single member of a vast space of all possible metabolic networks, which contains

2^{5906} distinct genotypes. We define the phenotype of a given metabolic genotype based on its viability on 50 distinct minimal environments that differ only in the carbon source (See Text S1c). We consider that a genotype is *viable* on a given carbon source, if it can produce all the essential biomass precursors from the given carbon source, and we use Flux Balance Analysis (FBA, See Text S1d) to determine viability [12–14]. We represent the phenotype of a given metabolic genotype as a binary vector of length 50. Each entry of this vector corresponds to a given carbon source, and it is equal to one if the genotype is viable on this carbon source, and zero otherwise.

(b) Reaction universe

The reaction universe is a set of metabolic reactions known to occur in some organism. For the construction of this universe, we used data from the LIGAND database [9,15]of the Kyoto Encyclopedia of Genes and Genomes [11,16]. Briefly, the LIGAND database, which is comprised of the REACTION and the COMPOUND databases, provides information on reactions, associated stoichiometric information, chemical compounds involved, and the Enzyme Classification (E.C.) identifier of each reaction. We used the REACTION and the COMPOUND databases to construct our universe of reactions, and excluded (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and, (iv) reactions involving complex metabolites without chemical information about their structure [8]. The published *E. coli* metabolic model (iAF1260) consists of 1397 non-transport reactions [12]. We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the unique (non-duplicate) reactions. This resulted in a universe of reactions consisting of 682 transport, 5906 non-transport reactions and 5030 metabolites.

(c) Chemical environments

We consider 50 minimal growth environments, each of which included oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe2+ and Fe3+), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese, zinc, and a specific carbon source Importantly, to represent different chemical environments, we vary the carbon source while keeping all other nutrients constant. We consider a metabolic network viable on a given carbon source, if it can synthesize all essential biochemical precursors when this carbon source is provided as the sole carbon source in a minimal medium.

We used 50 carbon sources for our analysis of randomly sampled metabolic networks, including the following 27 glycolytic carbon sources: D-Glucose, D-Glucose 6-phosphate, Trehalose,

Maltose, Lactose, D-Fructose 6-phosphate, D-Fructose, D-Mannose, D-Mannitol, D-Glucose 1-phosphate, D-Sorbitol, Maltotriose, D-Allose, D-Ribose, D-Xylose, D-Gluconate, 5-dehydro-D-Gluconate, L-Rhamnose, L-Fucose, L-Arabinose, L-Lyxose, D-Galactose, Melibiose, D-Galactonate, N-Acetyl-D-glucosamine, N-Acetyl-D-mannosamine, N-Acetylneuraminate.

In addition, we used the following 23 gluconeogenic carbon sources: Pyruvate, L-Alanine, L-Lactate, D-Alanine, D-Malate, Acetate, L-Serine, L-Malate, D-Serine, Glycine, Glycolate, L-Aspartate, Succinate, Fumarate, 2-Oxoglutarate, D-Galacturonate, D-Galactarate, D-Glucarate, L-Galactonate, D-Glucoronate.

And we used 3 nucleosides carbon sources: Adenosine, Deoxyadenosine, Inosine.

For the analysis of prokaryotic metabolic networks in the BiGG database, we used the following 137 carbon sources:
Acetaldehyde, Acetate, Acetoacetate, Adenine, Adenosine, Allantoin, Bicarbonate, Biotin, Butyrate (n-C4:0), Carbonic acid, Choline, Citrate, Cyanate, Cytidine, Cytosine, D-Alanine, D-Fructose, D-Galactarate, D-Galactonate, D-Galactose, D-Galacturonate, D-Glucarate, D-Gluconate, D-Glucosamine, D-Glucose, D-Glucose 6-phosphate, D-Glucuronate, D-Glyceraldehyde, D-Lactate, D-Mannitol, D-Mannose, D-Mannose 6-phosphate, D-Methionine, D-Ribose, D-Serine, D-Sorbitol, D-Xylose, Deoxyadenosine, Deoxycytidine, Deoxyguanosine, Deoxyinosine, Deoxyuridine, Dihydroxyacetone, Dimethyl sulfide, Dimethyl sulfoxide, Ethanol, Folate, Formate, Fumarate, Galactitol, Gamma-butyrobetaine, Glycerol, Glycerol 3-phosphate, Glycine, Glycine betaine, Glycolate, Guanine, Guanosine, Hexadecanoate (n-C16:0), Hypoxanthine, Indole, Inosine, L-Alanine, L-Arabinose, L-Arginine, L-Asparagine, L-Aspartate, L-Carnitine, L-Cysteine, L-Fucose, L-Fucose 1-phosphate , L-Glutamate, L-Glutamine, L-Histidine, L-Idonate, L-Isoleucine, L-Lactate, L-Leucine, L-Lysine, L-Malate, L-Methionine, L-Phenylalanine, L-Proline, L-Rhamnose, L-Serine, L-Threonine, L-Tryptophan, L-Tyrosine, L-Valine, L-tartrate, Lactose, Maltohexaose, Maltopentaose, Maltose, Maltotetraose, Maltotriose, Melibiose, Meso-2,6 Diaminoheptanedioate, Methanol, N-Acetyl-D-glucosamine, N-Acetyl-D-mannosamine, N, Acetylneuraminate NMN, Nicotinamide adenine dinucleotide, Octadecanoate (n-C18:0), Ornithine, Phenylpropanoate, Pimelate, Protoheme, Putrescine, Pyruvate, Riboflavin, Spermidine, Succinate, Sucrose, Taurine, Tetradecanoate (n-C14:0),Thiamin, Thymidine,Trehalose, Trimethylamine, Trimethylamine N-oxide, Uracil, Urea, Uridine, Xanthine, Xanthosine, AMP, (R)-Pantothenate, S)-Propane-1,2-diol, 1,5-Diaminopentane, 2-Dehydro-3-deoxy-D-gluconate, 2-Oxoglutarate, 3-(3-hydroxy-phenyl)propionate, 3-hydroxycinnamic acid.

(d) Flux balance analysis

Flux balance analysis (FBA) is a computational method that is widely used for the quantitative analysis and modeling of metabolic networks [13]. Based on the stoichiometric coefficients of the metabolites participating in the reactions of a given metabolic network, FBA predicts the metabolic flux through each reaction. Stoichiometric coefficients are stored in a stoichiometric matrix S, which is of dimension $m \times n$, where m and n, respectively, denote the number of metabolites and the number of reactions in a metabolic network. FBA constrains the flux through each reaction based on the assumption that a metabolic network is in a steady state where metabolite concentrations do not change, i.e., $Sv = 0$, where v is the vector of metabolic fluxes v_i through reaction i. The solutions of the equation $Sv = 0$, that is, the nullspace of matrix S, comprises all flux vectors that are allowable in steady state. The null space is further constrained by physicochemical information regarding the maximum and minimum possible flux through each reaction. FBA relies on an optimization procedure called linear programming to identify those flux vector(s) among the allowable ones that maximize an objective function Z. This task can be formulated as finding a flux vector v^* with the property

$$v^* = max_v\, Z(v) = max_v \{ c^T v \mid Sv = 0,\ a \leq v \leq b \},$$

where the vector c contains a set of scalar coefficients representing the maximization criterion, and each entry a_i and b_i of vectors a and b, respectively, indicates the minimally and maximally possible flux through reaction i. The vector c represents the proportions of each small biomass molecule in a cell's biomass. Therefore v^* maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass [14]. Here we use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and we consider a metabolic network viable if it can produce all essential biomass precursors. In a free-living bacterium like *E.coli,* there are approximately 60 such molecules including 20 amino acids, DNA, and RNA precursors, lipids and cofactors. We used the biomass composition of the *E. coli* metabolic model iAF1260 to define the vector c [12]. Moreover, we used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problem of FBA.

The major limitation of FBA is that it neglects regulatory constraints that can arise through suboptimal expression or regulation of enzymes. Newly horizontally transferred genes cannot easily establish regulatory interactions with their host genes, and it may thus take considerable adaptive evolution until they become expressed at a maximal or optimal level [17]. Such regulatory constraints would be especially important if we focused on quantitative predictions of biomass growth [18]. However, we use FBA solely for qualitative prediction of viability. This focus on qualitative phenotypes is biologically sensible. The reason is that many organisms grow slowly in their native environment [19–21], implying that regulation for maximal biomass production is far

from universal. Moreover, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments [18,22,23].

(e) Generation of random metabolic networks

We here employ a previously described *in silico* process which relies on Markov Chain Monte Carlo (MCMC) random walks to generate metabolic networks that comprise random sets of metabolic reactions that are viable on a given carbon source [8,24]. This procedure can produce metabolic networks that are sampled uniformly from the set of all metabolic networks viable on a given carbon source [8,24]. Briefly, in each step of such a random walk we perform a reaction swap, which is defined as altering a metabolic network by adding a randomly chosen reaction from the reaction universe, and then deleting a reaction randomly chosen from the set of reactions present in the metabolic network. If the reaction swap disrupts the metabolic network's viability on the given carbon source (as determined by FBA) we reject it, and perform another reaction swapping until we find a reaction swap that does not disrupt viability. This procedure also ensures that the total number of reactions remains constant. For the MCMC method to produce random samples of metabolic networks, it is essential to carry out enough reaction swaps to "erase" the random walker's similarity to the initial metabolic network. Previously, it has been shown that 3×10^3 reaction swaps are sufficient for this purpose [8,24]. Each of our random walks starts from *E. coli*'s metabolic network and performs 10^4 reaction swaps before storing the final metabolic network for further analysis. We used 10^4 independent random walks conducted in this way to create 10^4 random metabolic networks viable on glucose. We used the same procedure to generate 10^4 random metabolic networks viable on acetate.

(f) Generation of parental metabolic network pairs

Our analyses required us to recombine pairs of "parental" metabolic networks with particular features, such as (i) their genotypic distance (D), defined as the number of reactions differing between the parents, (ii) their phenotypic complexity ($||P||$), that is, the number of carbon sources on which they are viable, (iii) their phenotypic distance (ΔP), that is, the number of carbon sources on which only one but not the other member of a parental pair is viable, and (iv) their genotypic complexity ($||G||$, or metabolic network size), defined as the number of reactions in each metabolic network pair.

To identify parental metabolic networks with a given ΔP and $||P||$ we first selected, among all $\binom{10^4}{2}$ possible random metabolic network pairs that can be formed from 10^4 MCMC-sampled metabolic networks, those pairs that are viable on exactly $||P||$ carbon sources and that have a given ΔP. We then randomly chose from them a set of 1000 pairs for further analysis.

Less straightforward than identifying parental metabolic networks with a given ΔP and $||P||$ is to identify those with a given genotypic distance (D), because the random metabolic networks generated by MCMC sampling generally have genotypic distances sufficiently large ($D \approx 2000$) to be biologically unrealistic for modeling frequently recombining prokaryotic genomes. To create less diverse metabolic network pairs, we took two different MCMC random walk approaches that yielded similar results. The first revolves around a reaction-swapping random walk starting with a pair of randomly chosen metabolic networks from our sample of 10^4 sampled metabolic networks. In each step of this random walk, we subjected each parental metabolic network to a reaction swap, and we accepted each reaction swap if it (i) preserved the original phenotype, and (ii) did not increase the genotypic distance of the two metabolic networks after the swap, otherwise we rejected the reaction swap. We continued this procedure until the genotypic distance between the metabolic networks became equal to a desired distance D. This approach is very time-consuming. The second approach is much faster and uses a more biologically inspired mechanism to generate metabolic networks (see Text S2 [24,25]), but it also suffers from a technical limitation (Text S2), which is why we report mostly on the first approach.

Finally, to generate parental metabolic networks with a given number of reactions $||G||$ we started from a random viable metabolic network generated by MCMC sampling, as described in the Text S1e. All such metabolic networks have the same number of reactions as *E.coli* (2079). We then applied a sequence of reaction deletions that preserved viability on glucose (or acetate, depending on analysis) until we reached the desired $||G||$. Then, we sampled pairs of metabolic networks with a given D, ΔP and $||P||$ among the metabolic networks with $||G||$ reactions in the manner described above.

(g) Modeling recombination and mutation in metabolic networks

Prokaryotic genomes undergo recombination via horizontal gene transfer [26], whose incidence is large and greater than that of point mutations [27–29]. It changes the organization and gene content of genomes on short evolutionary time scales [26,30,31], and can involve very distantly related organisms [32,33]. Various mechanisms of horizontal gene transfer add genes unidirectionally from a donor to a recipient, but incorporating such genes into the recipient genome relies on recombination [26]. The genomes of many prokaryotes frequently undergo homologous recombination, that is, a reciprocal exchange of DNA segments between DNA sequences [34]. Because such recombination can also delete genes, and because of a general deletion bias in prokaryotic genomes [35], prokaryotic recombination involves gene loss as well as gene gain. What is more, the majority of newly acquired genes obtained via horizontal gene transfer reside in the genome only for short amounts of time [36]. Motivated by these observations, we here model prokaryotic recombination as a process where the transfer of reactions from a donor to a recipient metabolic network is compensated by deletion of other reactions from the recipient.

To model recombination for each parental metabolic network pair, we generated 1000 recombinant offspring by (i) adding to the recipient metabolic network a given number *n/2* of randomly chosen reactions that were present in the donor and absent in the recipient, followed by (ii) deleting *n/2* reactions randomly chosen from the recipient. Thus, the total number of reactions changed by a recombination event in the recipient is equal to *n*. For reasons of computational feasibility, we analyzed only recombinant pairs where the probability that a recombination event preserves viability exceeded 10^{-3}. Text S3 and figure S1 show that this is the case for values of *n* up to *60,* which is why we chose *n=60* as the highest amount of reaction changes during a recombination event. Empirical observations also suggest that this number of reactions would not be unrealistically large, because horizontal gene transfer can affect long DNA regions[44]. Transferred material that is integrated into the host genome by recombination can constitute stretches of non-coding DNA, fragments of genes [37,38], entire genes [39], multiple adjacent genes [40,41], operons, transposable chromosomal elements, plasmids, as well as other naturally occurring extrachromosomal elements [42]. The length of contiguous transferred stretches may range from a few nucleotides [43] to more than 3 Mbp [44], i.e., some two thirds of the length of the *E.coli* genome, which encodes more than 1300 reactions. In addition, some megabase-scale horizontally transferred genes can become incorporated into a chromosome in the form of hundreds of smaller fragments [45].

To implement an amount of random mutational metabolic change that is comparable to the same amount of recombinational change, for a given number of altered reactions (*n*) we created a "mutational" offspring of a metabolic network by adding *n*/2 randomly chosen reactions from the reaction universe, and deleting *n/2* randomly chosen reactions among the set of reactions present in the metabolic network. Note the key difference between mutation and recombination: In recombination the *n/2* reactions that are added to the recombinant offspring are chosen randomly from another viable metabolic network (the donor), whereas in mutation they are taken from the whole reaction universe.

(h) Genomic recombination in prokaryotic metabolic networks from the BiGG database

We validated our observations based on randomly sampled viable metabolic networks by considering the genome-scale metabolic networks of 61 bacterial species available at the BiGG database [46], using the R-package Sybil [47]. For this analysis, we generated a reduced universe of reactions comprised of the union of the sets of reactions present in the 61 metabolic networks. This universe altogether contains 3404 internal reactions, 3156 transport reactions, and a different biomass reaction for each organism. As potential carbon sources, we used all 137 carbon-containing metabolites that occurred as metabolites external to at least one organism in the database, and thus assigned a phenotype vector of length 137 to each metabolic network using FBA.

To model recombination among the metabolic networks of these 61 organisms, we used one main approach, which incorporates information about the linkage of the genes encoding metabolic reactions. To this end, we used the gene-reaction association rules defined in the BiGG database for each organism (in MAT files, grRules) [46], and ordered the genes in each organism based on their genomic position, as obtained from the RefSeq microbial genome database [48].

For a specific recombination event between a donor and a recipient organism, we first chose at random a stretch of DNA from the donor organism that contains a given number of metabolic genes. To generate a recombinant offspring we added this stretch of DNA to the recipient, and subsequently deleted a randomly selected stretch of DNA from the recipient genome. We translated the added and deleted genes into reactions based on the gene-reaction rules for the donor and recipient organism. We set the number of genes in every donor DNA stretch such that on average (among all recombination events between all metabolic network pairs) a given number of n reactions are added to the recipient metabolic network, and an equal number n of reactions are deleted from it. Because gene-reaction associations are not generally one-to-one and can be very complicated, and because most of the reactions that are encoded in a given stretch of DNA may already be present in the recipient metabolic network, the number of metabolic genes in donor DNA required for adding n reactions will be higher than n (usually $\approx 2n$). In contrast, we found that including $\approx 0.9n$ metabolic genes into a DNA stretch to be deleted from the recipient genome usually sufficed to eliminate n reactions from the recipient metabolic network, because deletion of a single metabolic gene often causes elimination of multiple reactions

In a second approach for recombining prokaryotic genomes, we neglected linkage between metabolic genes and added or deleted reactions randomly, just as we had done for randomly sampled viable metabolic networks, irrespective of the genomic position of metabolic genes encoding these reactions.

Text S2:

An alternative MCMC approach to generate parental metabolic networks with a given genotypic distance (*D*)

In addition to our first and main approach (see text S1f) for creating metabolic networks with a given genotypic distance D, we also pursued a second approach. This second approach starts from a parental metabolic network M_1 and generates a recombination partner M_2 through a sequence of MCMC random walks, which preserves the phenotype of M_1 but increases the genotypic distance to M_2, until a desired D between M_1 and M_2 is reached. Because this method resembles the divergence of species from a common ancestor, it is biologically motivated, but it has a technical limitation that is associated with inactive (blocked) reactions – reactions having zero metabolic flux for stoichiometric reasons [24,25]. Probably due to the shorter MCMC random walks in this second

approach, a greater percentage of the *D* reactions that are not shared between two metabolic networks are blocked reactions in the second approach (90.09%) compared to the first one (66.78% of reactions). The innovation potential of inactive reactions is almost negligible in comparison to active reactions (Figure S2), and the fraction of innovative offspring is thus considerably (almost an order of magnitude) lower for the second approach than for the first approach.

To render results from the two approaches comparable, one can adjust the ratio of inactive reactions in the *D* non-shared reactions between metabolic network pairs, as illustrated in the following example. Let us assume a given parental metabolic network pair obtained with the second method has *D=1000* non-shared reactions, and since ≈90% of these reactions are inactive (blocked), only around 100 of them will be active. To make the ratio of active to inactive reactions among these non-shared reactions equal to the 66.78% (≈2/3) that are characteristic of parental metabolic network pairs obtained with the first method, one would require almost 200 inactive reactions (200/(100+200)= 2/3). Thus, whenever one wants to transfer a given number of reactions (*n*) from donor to recipient, one can first select 200 inactive reactions from the 900 inactive reactions and then select the *n* reactions to be recombined from a set that includes these 200 inactive reaction and 100 active reactions. This ensures that a comparable proportion of active reactions are transferred from donor to recipient in the two approaches. After this adjustment, the two approaches yield virtually identical observations (Compare figure1 with figure S3). However, the manipulations required in the second approach make it less useful. We therefore chose to rely on the first approach throughout this study.

Text S3:

Robustness of genome-scale metabolic networks decreases exponentially with increasing the number of deleted reactions

Recombination that involves both the addition and deletion of reactions has the potential to create inviable recombinant offspring. The greater the number of reactions that are deleted in a recombination event, the greater will be this fraction of inviable offspring. Before embarking on a systematic analysis of recombination's effects, we needed to find out how large the number of reactions deleted in a recombination event (*n/2*) can become, before the number of viable offspring becomes too small for computational analysis. To this end, we generated 1000 random genome-scale metabolic networks that we required to be viable only on glucose as the sole carbon source. For each of these randomly sampled viable metabolic networks and for each value of *n* between one and sixty, we created 1000 offspring in which we deleted *n/2* randomly chosen reactions. Figure S1 shows a box plot of the fraction of metabolic networks that remain viable on glucose as the sole carbon source after this procedure.

The fraction of viable metabolic networks declines exponentially with the number of deleted reactions. For $(n/2)>30$ the fraction of metabolic networks that retain viability becomes very low, e.g., it declines below 0.001 for viability on glucose, such that fewer than one of 1000 offspring would be viable on glucose. At numbers beyond $(n/2)>30$, the number of recombination events needed to create any viable metabolic networks becomes computationally prohibitive. For this reason, we chose $n=60$ as the highest value of n for our recombination analysis.

Text S4:

The rate of recombination between bacterial species decreases exponentially by increasing metabolic distance

We wanted to obtain a crude estimate of the relationship between the metabolic distance of two bacterial species and the likelihood that such species undergo a successful homologous recombination event. To estimate this relationship, we pursued a three-step procedure.

In the first step, we estimated the DNA-based genotypic distance between two bacterial species whose metabolic networks differ by a given number of reactions. To this end, we used curated metabolic networks from 51 bacterial species, which had been obtained through state-of-the art techniques for genome annotation, generation of biomass reactions, reaction network assembly, and thermodynamic analysis of reaction reversibility [49]. We define the normalized metabolic genotype distance d of two prokaryotes as the number D of reactions differing between their metabolic networks, divided by the total number of reactions present in at least one of the two metabolic networks and computed this distance for all pairs of the 51 metabolic networks. On average, a relative metabolic distance of $d=0.1$ corresponds to an absolute difference of $D\cong150$ in reaction number, but we note that the relationship between d and D depends on the total number of reactions in each metabolic network.

We then aimed to relate metabolic divergence to DNA sequence divergence between these species. To this end, we used the housekeeping gene rpoB, which encodes the β-subunit of RNA polymerase. We obtained the rpoB coding sequences for these 51 species from NCBI (http://www.ncbi.nlm.nih.gov), and aligned them with the PAL2NAL web server, which provides robust alignment of DNA sequences based on the corresponding protein sequences [50]. We then computed all pairwise Hamming distances from the aligned rpoB sequence alignment for these 51 species, normalized these quantities to the interval (0,1), and used them as our measure of sequence divergence.

We note that even species with modest sequence divergence can have considerable metabolic distance. For example, the species pair *Buchnera aphidicola* and *Yersinia pestis* have an *rpoB* DNA sequence distance of 0.29, but a metabolic distance d of 0.65, which corresponds to an absolute

difference of 1004 reactions. Examples of moderately high metabolic divergence exist even from strains of the same organisms, such as *Streptococcus pneumoniae* TIGR4 and R6, which differ in 64 reactions or a fraction $d=0.068$ of their metabolic networks. At greater sequence distances of 0.45, metabolic distances reach values up to $d=0.69$ (e.g., *Yersinia pestis* and *Rickettsia prowazekki* have *rpoB* DNA sequence divergence 0.45 and a metabolic distance of $d=0.684$, corresponding to 1066 reaction differences.) Metabolic distance and sequence divergence are significantly correlated (Pearson's $r=0.60$, $P<10^{-40}$), and a linear regression analysis (red line in figure S4a) yields a regression coefficient of 0.82 with an intercept of 0.1. We use this regression analysis to translate metabolic distance into sequence divergence and vice versa.

In the second step, we took advantage of experimental data on the exponential relationship between the likelihood of a successful recombination event and (rpoB-based) sequence divergence between recombining species [33,51]. Specifically, we used such data for 19 species pairs in the genera Bacillus and Streptococcus. Figure S4b shows that the logarithm of the relative recombination rate decreases linearly with increasing sequence divergence between the donor and recipient species. A linear regression analysis (black line in the figure) yields a regression coefficient of -18.40 with an intercept of 0.11.

In the third step, we integrated data from step one and two to relate metabolic distance to the likelihood of a successful recombination event (Figure S4c). The figure shows that the logarithm of the relative recombination rate linearly decreases with increasing metabolic distance between the donor and recipient species. A linear regression analysis (red line in the figure) yields a regression coefficient of -22.57 with an intercept of 0.62. In sum, sequence and recombination data suggests that the likelihood of a successful recombination event between two species would decrease exponentially with their metabolic distance. This also holds if we exclude endosymbiotic or host-associated pathogens from our analysis.

Importantly, we note that metabolic distance will not be the only determinant of successful recombination between bacteria of different species. Part of the reason is that only a minority of genes in any bacterial genome are typically involved in metabolic network (e.g., 31% in *E.coli*). In addition, other incompatibilities, such as those between restriction-methylation systems [52] or DNA repair mechanisms [53] may hinder recombination. Our analysis merely goes to show that the minimal recombination distances of $D=100$ we use are not unrealistically low. Many bacteria that would successfully recombine in the wild have greater metabolic distances (Figure S4c).

Text S5:

Phenotypically more diverse parental metabolic networks are more likely to generate metabolically innovative offspring

We asked whether the phenotypic diversity of recombining parents influences the incidence of innovative offspring. On the one hand, recombining parents viable on the same combination of carbon sources might create a greater fraction of viable offspring, which might also increase the incidence of offspring with novel metabolic abilities. On the other hand, recombining parents viable on different combinations of carbon sources might produce recombinant offspring with a greater number of *novel* reaction combinations, and thus a greater number of metabolic innovations.

To find out whether one of these hypotheses is correct, we created pairs of metabolic networks at a fixed genotypic distance ($D=100$), but with different metabolic phenotypes P_1 and P_2 and with identical phenotypic distances (ΔP), that is, identical number of carbon sources on which one parent is viable but the other isn't, or vice versa. To prevent confounding our analysis by the number of carbon sources $||P||$ on which a metabolic network is viable, we kept $||P||$ constant and required that each parent was viable on exactly 10 carbon sources. (In other words, all metabolic networks in this analysis are viable on glucose and on nine other carbon sources.) We then varied ΔP in four steps between 0 and 16, created 1000 metabolic network pairs for each value of ΔP, and from each pair we created 1000 recombinant offspring in which n reactions were altered through recombination. We then determined for each offspring whether it was viable on any carbon source that neither of the parents were viable on. Figure 2c (main text) shows that regardless of the number n of altered reactions, the fraction of innovative offspring (f_{innov}) increases with the phenotypic distance ΔP among parents.

The increase of innovation with parental phenotypic diversity cannot just be explained by a greater fraction of viable offspring, because parental phenotypic diversity does not influence this fraction (Figure S10a and S10b). In contrast, as ΔP increases, so does the fraction of reactions with $I_{SE} > 0.5$ that can potentially be transferred from donor to recipient (Figure 2d (main text)), once again highlighting the role of this process in innovation. Parental phenotypic diversity ΔP does have no impact on the number of carbon sources on which innovative offspring gain viability. Specifically, we observed that innovative offspring typically gains viability on two to three additional carbon sources, and shows an average phenotypic distance between four and five, regardless of whether it arose through recombination or mutation, and independent of parental genotypic or phenotypic features.

We complemented these analyses by focusing on an alternative way of defining phenotypic heterogeneity that is based on viability on two specific classes of carbon sources, namely those

involved primarily in glycolysis, and those involved primarily in gluconeogenesis (See Text S1c). We found that offspring of parents viable on different classes of carbon sources display a greater incidence of innovation, compared to offspring of parents that are viable on the same class of carbon sources (Text S6).

Text S6:

Parental metabolic networks viable on different classes of carbon sources are more likely to generate innovative offspring than parents that are viable on the same classes of carbon sources

In this analysis, we focused on two specific classes of carbon sources, namely those involved primarily in glycolysis, and those involved primarily in gluconeogenesis (Text S1c). In a previous contribution, we had shown that metabolic networks required to be viable on one glycolytic (gluconeogenic) carbon source tended to be viable also on other glycolytic (gluconeogenic) carbon sources [54]. We wanted to find out whether parental viability on either glycolytic, gluconeogenesis, or both kinds of carbon sources influenced the incidence of novel metabolic traits in the offspring. To this end, we created 1000 pairs of donor – recipient metabolic networks (genotype distance $D=100$) with each of the following properties (i) both parents are viable on five glycolytic carbon sources, (ii) both parents are viable on five gluconeogenic carbon sources, (iii) all donor metabolic networks are viable on five gluconeogenic carbon sources, and all recipient metabolic networks are viable on five glycolytic carbon sources, and (iv) all donor metabolic networks are viable on five glycolytic carbon sources, and all recipient metabolic networks are viable on five gluconeogenic carbon sources. To exclude parental phenotypic diversity as a confounding factor, we ensured that it had a constant value of $\Delta P=10$ for all parents in all three categories. Aside from these constraints, we chose glycolytic and gluconeogenic carbon sources at random.

For each pair of metabolic networks we created 1000 offspring with a fixed number of altered reactions, and found that recombinants of parents viable on different kinds of carbon sources (i.e. gluconeogenic-glycolytic) display a greater incidence of innovation (Figure S11a). This greater incidence of innovation cannot solely be explained by a greater fraction of viable offspring, because parental viability on different classes of carbon sources does not influence the fraction of viable offspring (Figure S11b). Thus, we conclude that phenotypically more heterogeneous parental metabolic networks are more likely to generate innovative recombinants.

Text S7:

Phenotypically less complex parental metabolic networks are more likely to generate metabolically innovative offspring

We also investigated the impact of phenotypic complexity on metabolic innovation. We define the complexity of a phenotype P as the number $||P||$ of carbon sources on which it is viable. For this analysis, we generated parental metabolic networks with the same genotypic distance $D=100$ but with varying phenotypic complexity. In addition, we required that both metabolic networks in a pair are viable not only on the same number of carbon sources, but also on the exact same carbon sources. Specifically, we analyzed 1000 pairs of random parental metabolic networks viable on $||P||=1$, 5, or 10 carbon sources. For each of the 1000 pairs at each value of $||P||$, we created 1000 recombinant offspring with n altered reactions. Figure 2e (main text) shows that the fraction of innovative offspring (f_{innov}) decreases with increasing phenotypic complexity. The more carbon sources a metabolic network is viable on, the smaller the likelihood that recombination creates viability on further carbon sources. This difference is not simply caused by a decrease in the fraction of viable offspring with increasing phenotypic complexity (Figures S12a and S12b). Also, $||P||$ does not impact the number of additional carbon sources that an innovative offspring gains viability on. The fraction of exchangeable reactions with $I_{SE} > 0.5$ (f_{super}) decreases with increasing phenotypic complexity (Figure 2f (main text)).

We also wished to analyze the effect of phenotypic complexity on metabolic innovation for metabolic networks viable on more than 10 carbon sources (i.e., $||P||=20$, 30, 40). However, creating 1000 metabolic network pairs with these values of $||P||$ that were viable on the exact same combination of carbon sources was computationally infeasible. We thus created 1000 metabolic network pairs whose phenotype vectors differed in a fixed number of 10 non-zero entries. For example, in such a pair with $||P||=30$, both members would be viable on the same 25 carbon sources. In addition one member would be viable on five carbon sources that the other one is not viable on, and vice versa. Furthermore, we required a fixed genotypic distance $D=100$ for all metabolic network pairs in this analysis. For each value of $||P||$ subject to these constraints, we created 1000 recombinant offspring with n altered reactions for each of the 1000 parental metabolic network pairs. Consistent with data from figure 2e (main text), the fraction of innovative offspring (f_{innov}) decreases with increasing phenotypic complexity (Figure S13a). Figures S13b and S13c show that this difference is not simply caused by a decrease in the fraction of viable offspring with increasing phenotypic complexity. Figure S13d shows that fraction of reactions with superessentiality higher than 0.5 (f_{super}) decreases with increasing phenotypic complexity.

Text S8:

Larger parental metabolic networks are recombinationally more robust and so more likely to generate metabolically innovative offspring

We define the genotypic complexity of a metabolic network as the number of reactions ($||G||$) present in this metabolic network. We wanted to find out whether it affects recombinational robustness and the incidence of novel phenotypes among recombinant offspring. To this end, we analyzed metabolic networks with sizes that vary between $||G||=1500$ and $||G||=2000$ reactions. Specifically, we created 1000 random viable donor recipient pairs with constant genotype distance $D=100$ for each size class, where we required the parental metabolic networks to be viable only on glucose. We then created from each parental pair 1000 recombinant offspring with a specific number n of altered reactions.

Figure S14a shows that recombinational robustness increases with increasing genotypic complexity of the parents. For example, at $n=10$ recombined reactions the fraction of viable offspring is three times higher for parental metabolic networks with $||G||=2000$ reactions than for parental metabolic networks with $||G||=1500$ reactions (0.3 vs. 0.1, Figure S14a). We observe the same result, when we use parental metabolic networks viable on acetate for this analysis (Figure S14b). Moreover, we observed that the fraction of innovative offspring (f_{innov}) also increases with increasing metabolic network size $||G||$ (Figure S14c). Again this result does not change if parental metabolic networks are viable on acetate instead of on glucose (Figure S14d).

In sum, unlike other quantities such as genotypic distance and phenotypic diversity of parents, which do not impact recombinational robustness, and thus influence innovation directly, parental genotypic complexity ($||G||$) increases recombinational robustness, and can thus enhance innovation indirectly by increasing robustness.

Text S9:

Effects of genotypic and phenotypic features of prokaryotic parental metabolic networks on recombinational robustness and innovation

Unlike our analyses of random viable metabolic networks, where we were able to control genotypic parameters such as the number of reactions, and phenotypic parameters such as phenotypic complexity by sampling genotype space appropriately, these parameters are fixed properties of the 61 specific prokaryotic metabolic networks we analyzed. Moreover, when analyzing random viable networks we could sample metabolic network pairs that varied in only one parameter, which is not

possible for prokaryotic metabolic networks. However, to control relevant parameters to some extent, we took the following steps.

First, to prevent size variation in metabolic networks from confounding our analysis, we observed that the majority (47 of 61) of prokaryotic metabolic networks show a narrow size range between 1250 and 1350 internal reactions (Figure S16a), and focused our analysis on these metabolic networks. (278 among all $\binom{47}{2}$ possible pairs of these metabolic network pairs have at least one offspring that is viable on a new carbon source.)

Second, we observed that the distribution of parental genotypic distance (D), phenotypic distance (ΔP), and phenotypic complexity ($||P||$) is distinctly bimodal for these 278 parental metabolic networks (Figures S16b, S16c, and S16d). No parental metabolic network has intermediate values for any of these parameters, such that metabolic networks can be subdivided into "high" and "low" categories for each parameter. Moreover, metabolic networks with high ΔP have low $||P||$ (Figure S16e). Based on these observations, we subdivided parental metabolic network pairs into 4 categories: i) high D and low ΔP (high $||P||$), ii) low D and low ΔP (high $||P||$), iii) high D and high ΔP (low $||P||$), iv) low D and high ΔP (low $||P||$). The number of parental metabolic networks in categories (i) through (iv) was 96, 106, 12, and 64, respectively. Recombinational robustness differs little among metabolic networks in these four categories (Figures S17b and S17c), and so these parameters do not strongly influence robustness, which is consistent with our observations from random viable metabolic networks. In contrast, the fraction of innovative offspring of parental metabolic networks in the fourth category (with low D and high ΔP (low $||P||$)) is highest, and it is lowest for metabolic networks in the first category (with high D and low ΔP (high $||P||$)) (Figures 3c and S17a). This is again consistent with our observations from random viable metabolic networks, where parents with low genotypic distance, high phenotypic distance, and low phenotypic complexity are more likely to generate innovative offspring.

Text S10:

Superessential reactions can explain the effect of parental genotypic and phenotypic diversity and complexity on metabolic innovation.

Superessential reactions that are involved in recombination events can explain a series of patterns in our data. The first is that a given number of reaction changes can elicit more metabolic innovation when caused by recombination rather than by mutation. While recombination adds reactions to a recipient that already occur in a (viable) donor, random mutations add reactions unrelated to the metabolic network of the donor. Because this reaction universe contains fewer highly super-

essential reactions than any donor, adding reactions from it is less likely to yield innovations (electronic supplementary material, figure S18). Put differently, when recombination introduces new metabolic reactions into an organism, it preferentially introduces reactions that have been "pretested" by evolution, because they form part of a related viable genotype. In contrast, mutations may introduce reactions that are incompatible with this genotypic background, in the sense that they cannot interact productively with it. This observation is consistent with observations from other systems, such as proteins [55,56] and model gene regulatory networks [57,58].

Super-essential reactions can also help explain that the incidence of metabolic innovation rises with the number of transferred reactions (Figure 1a, main text). As we showed in the main text, addition of a single reaction is usually sufficient to cause metabolic innovation. By increasing the number of transferred reactions, the probability increases that at least one highly superessential reaction is transferred, and so the incidence of metabolic innovation increases.

In addition, the transfer of superessential reactions can help explain that increasing genotypic distance between donor and recipient decreases the incidence of metabolic innovation (figure 2a, main text). Since the number of highly superessential reactions is limited [59], increasing the genotypic distance between donor and recipient decreases the fraction of such reactions that are not already in the recipient. In consequence, the incidence of innovative offspring decreases as well.

Superessential reactions can also help explain why the incidence of innovation increases with increasing parental phenotypic diversity ΔP – an increasing number of carbon sources on which one but not the other parent is viable. Any phenotypic difference between parents must be caused by the set of D reactions that are not shared between the parents. As ΔP increases, an increasing number of these non-shared reactions would be involved in viability on at least one of the carbon sources on which the parents are viable, and such reactions tend to have a higher super-essentiality index [59]. These are also the reactions that will lead to innovation when affected by a recombination event (figure S7). Therefore, parents with higher ΔP are expected to have a higher fraction of exchangeable reactions with high super essentiality index (Figure 2d in main text), and consequently higher fraction of innovative offspring (Figure 2c in main text).

Finally, with increasing phenotypic complexity ||P|| – the number of carbon sources on which a metabolic network is viable – of parental metabolic networks with the same phenotype, the incidence of innovation by recombination decreases. To explain this pattern, consider two genotypically distinct metabolic networks with the same phenotype. Their non-shared reactions are less likely to be essential for viability than their shared reactions, and so the superessentiality index of the non-shared reactions is expected to be low. As ||P|| increases, the fraction of non-shared reactions with high superessentiality index is expected to decrease further, exactly as we observed (Figures 2f (main text) and S13d), which leads to a lower incidence of innovation (Figures 2e (main text) and S13a).

References:

1. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. 2001 In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–30. (doi:10.1038/84379)

2. Edwards, J. S. & Palsson, B. O. 1999 Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–6.

3. Lewis, N. E., Nagarajan, H. & Palsson, B. O. 2012 Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305. (doi:10.1038/nrmicro2737)

4. Feist, A. M. & Palsson, B. Ø. 2008 The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* **26**, 659–67. (doi:10.1038/nbt1401)

5. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. 2009 Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320. (doi:10.1038/msb.2009.77)

6. McCloskey, D., Palsson, B. Ø. & Feist, A. M. 2013 Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9**, 661. (doi:10.1038/msb.2013.18)

7. Fondi, M. & Liò, P. 2015 Genome-scale metabolic network reconstruction. *Methods Mol. Biol.* **1231**, 233–56. (doi:10.1007/978-1-4939-1720-4_15)

8. Matias Rodrigues, J. F. & Wagner, A. 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5**, e1000613. (doi:10.1371/journal.pcbi.1000613)

9. Goto, S., Nishioka, T. & Kanehisa, M. 2000 LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28**, 380–2.

10. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. 2010 KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**, D355–60. (doi:10.1093/nar/gkp896)

11. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. 2006 From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–7. (doi:10.1093/nar/gkj102)

12. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. & Palsson, B. Ø. 2007 A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121. (doi:10.1038/msb4100155)

13. Orth, J. D., Thiele, I. & Palsson, B. Ø. 2010 What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8. (doi:10.1038/nbt.1614)

14. Kauffman, K. J., Prakash, P. & Edwards, J. S. 2003 Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491–6.

15. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. 2002 LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30**, 402–4.

16. Kanehisa, M. & Goto, S. 2000 KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

17. Lercher, M. J. & Pál, C. 2008 Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* **25**, 559–67. (doi:10.1093/molbev/msm283)

18. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. 2002 Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186–9. (doi:10.1038/nature01149)

19. Vieira-Silva, S. & Rocha, E. P. C. 2010 The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808. (doi:10.1371/journal.pgen.1000808)

20. Cox, R. A. 2004 Quantitative relationships for specific growth rates and macromolecular compositions of Mycobacterium tuberculosis, Streptomyces coelicolor A3(2) and Escherichia coli B/r: an integrative theoretical approach. *Microbiology* **150**, 1413–26.

21. Kirschner, D. & Marino, S. 2005 Mycobacterium tuberculosis as viewed through a computer. *Trends Microbiol.* **13**, 206–11. (doi:10.1016/j.tim.2005.03.005)

22. Fong, S. S. & Palsson, B. Ø. 2004 Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.* **36**, 1056–8. (doi:10.1038/ng1432)

23. Fong, S. S., Marciniak, J. Y. & Palsson, B. O. 2003 Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* **185**, 6400–6408. (doi:10.1128/JB.185.21.6400-6408.2003)

24. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. 2010 Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4**, 30. (doi:10.1186/1752-0509-4-30)

25. Burgard, A. P., Nikolaev, E. V, Schilling, C. H. & Maranas, C. D. 2004 Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14**, 301–12. (doi:10.1101/gr.1926504)

26. Thomas, C. M. & Nielsen, K. M. 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–21. (doi:10.1038/nrmicro1234)

27. Guttman, D. S. & Dykhuizen, D. E. 1994 Clonal divergence in Escherichia coli as a result of recombination, not mutation. *Science* **266**, 1380–3.

28. Feil, E. J. et al. 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 182–7. (doi:10.1073/pnas.98.1.182)

29. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. 2005 Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol. Evol.* **22**, 2354–61. (doi:10.1093/molbev/msi233)

30. Ochman, H., Lawrence, J. G. & Groisman, E. A. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304. (doi:10.1038/35012500)

31. Pál, C., Papp, B. & Lercher, M. J. 2005 Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372–5. (doi:10.1038/ng1686)

32. Fraser, C., Hanage, W. P. & Spratt, B. G. 2007 Recombination and the nature of bacterial speciation. *Science* **315**, 476–80. (doi:10.1126/science.1127573)

33. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. 2000 Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J. Bacteriol.* **182**, 1016–23.

34. Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. & Rehrauer, W. M. 1994 Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.* **58**, 401–65.

35. Mira, A., Ochman, H. & Moran, N. A. 2001 Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–96.

36. Kuo, C.-H. & Ochman, H. 2009 The fate of new bacterial genes. *FEMS Microbiol. Rev.* **33**, 38–43. (doi:10.1111/j.1574-6976.2008.00140.x)

37. Bork, P. & Doolittle, R. F. 1992 Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci.* **89**, 8990–8994. (doi:10.1073/pnas.89.19.8990)

38. Inagaki, Y., Susko, E. & Roger, A. J. 2006 Recombination between elongation factor 1 genes from distantly related archaeal lineages. *Proc. Natl. Acad. Sci.* **103**, 4528–4533. (doi:10.1073/pnas.0600744103)

39. Hartl, D. L., Lozovskaya, E. R. & Lawrence, J. G. 1992 Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica* **86**, 47–53.

40. Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K. & Nagashima, K. V 2001 Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J. Mol. Evol.* **52**, 333–41. (doi:10.1007/s002390010163)

41. Omelchenko, M. V, Makarova, K. S., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V 2003 Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* **4**, R55. (doi:10.1186/gb-2003-4-9-r55)

42. Chan, C. X., Beiko, R. G., Darling, A. E. & Ragan, M. A. 2010 Lateral Transfer of Genes and Gene Fragments in Prokaryotes. *Genome Biol. Evol.* **1**, 429–438. (doi:10.1093/gbe/evp044)

43. Denamur, E. et al. 2000 Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* **103**, 711–21.

44. Lin, C. H., Bourque, G. & Tan, P. 2008 A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol. Biol. Evol.* **25**, 549–58. (doi:10.1093/molbev/msm282)

45. Didelot, X., Achtman, M., Parkhill, J., Thomson, N. R. & Falush, D. 2007 A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res.* **17**, 61–8. (doi:10.1101/gr.5512906)

46. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. 2015 BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* , gkv1049–. (doi:10.1093/nar/gkv1049)

47. Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. 2013 Sybil--efficient constraint-based modelling in R. *BMC Syst. Biol.* **7**, 125. (doi:10.1186/1752-0509-7-125)

48. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. 2014 RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* **42**, D553–9. (doi:10.1093/nar/gkt1274)

49. Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. & Stevens, R. L. 2010 High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–82. (doi:10.1038/nbt.1672)

50. Suyama, M., Torrents, D. & Bork, P. 2006 PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12. (doi:10.1093/nar/gkl315)

51. Zawadzki, P., Roberts, M. S. & Cohan, F. M. 1995 The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics* **140**, 917–32.

52. Jeltsch, A. 2003 Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* **317**, 13–6.

53. Zahrt, T. C. & Maloy, S. 1997 Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between Salmonella typhimurium and Salmonella typhi. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 9786–91.

54. Barve, A. & Wagner, A. 2013 A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203–6. (doi:10.1038/nature12301)

55. Cui, Y., Wong, W. H., Bornberg-Bauer, E. & Chan, H. S. 2002 Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 809–14. (doi:10.1073/pnas.022240299)

56. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. 2005 On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5380–5. (doi:10.1073/pnas.0500729102)

57. Martin, O. C. & Wagner, A. 2009 Effects of recombination on complex regulatory circuits. *Genetics* **183**, 673–84, 1SI–8SI. (doi:10.1534/genetics.109.104174)

58. Wagner, A. 2011 The low cost of recombination in creating novel phenotypes: Recombination can create new phenotypes while disrupting well-adapted phenotypes much less than mutation. *Bioessays* **33**, 636–46. (doi:10.1002/bies.201100027)

59. Barve, A., Rodrigues, J. F. M. & Wagner, A. 2012 Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1121–30. (doi:10.1073/pnas.1113065109)

Supplementary figures:

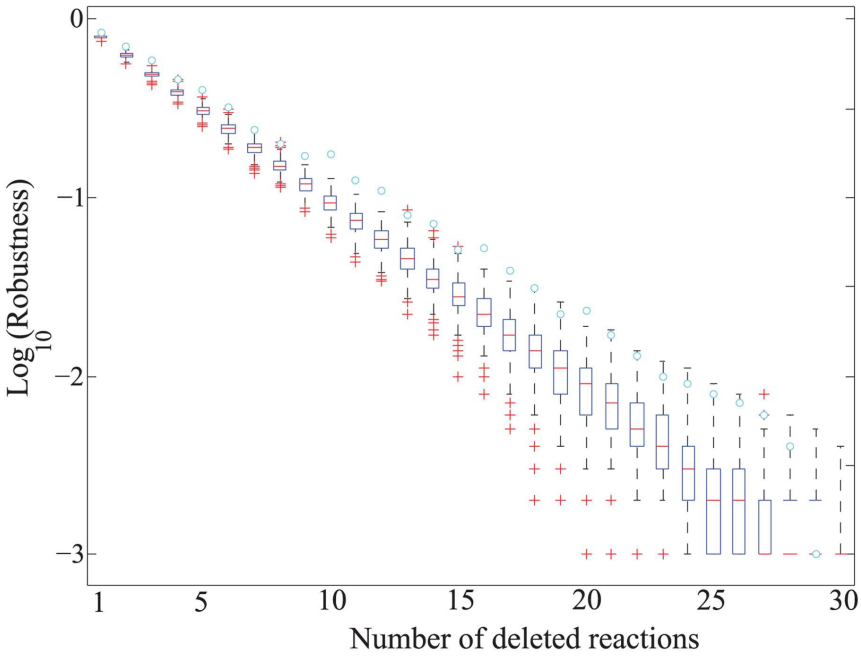


Figure S1: Distribution of the fraction of randomly sampled viable metabolic networks (boxes) that retain viability on glucose (*y*-axis, note the logarithmic scale) as compared with that of *E. coli* (cyan circles) after deleting a given number of reactions (*x*-axis). All boxes span the 25-th to 75-th percentile. Horizontal bars in a box indicate the median, and whiskers indicate maxima and minima. Red asterisks indicate outliers.

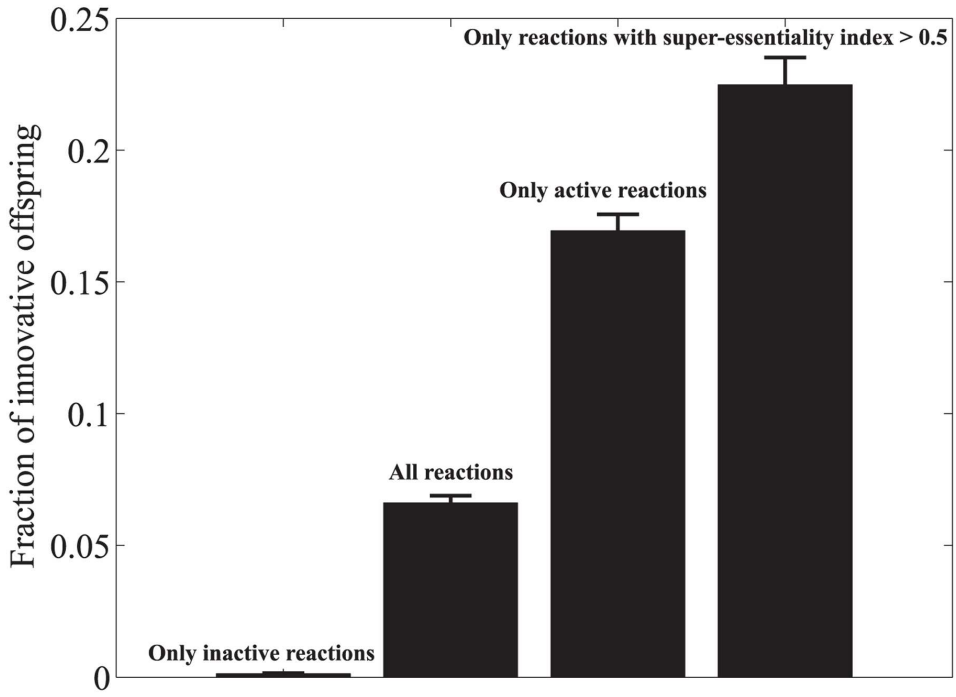


Figure S2: Mean (bar), and standard error (vertical line) of the fraction of innovative offspring (f_{innov}) generated from 1000 random parental metabolic networks viable only on glucose with a fixed genotype distance *D=100*, by adding 5 randomly chosen i) inactive (blocked), ii) active reactions, iii) highly superessential, and iv) mixed (including all types) reactions from a donor metabolic network to the recipient, followed by deleting 5 randomly chosen reactions from the recipient metabolic network.

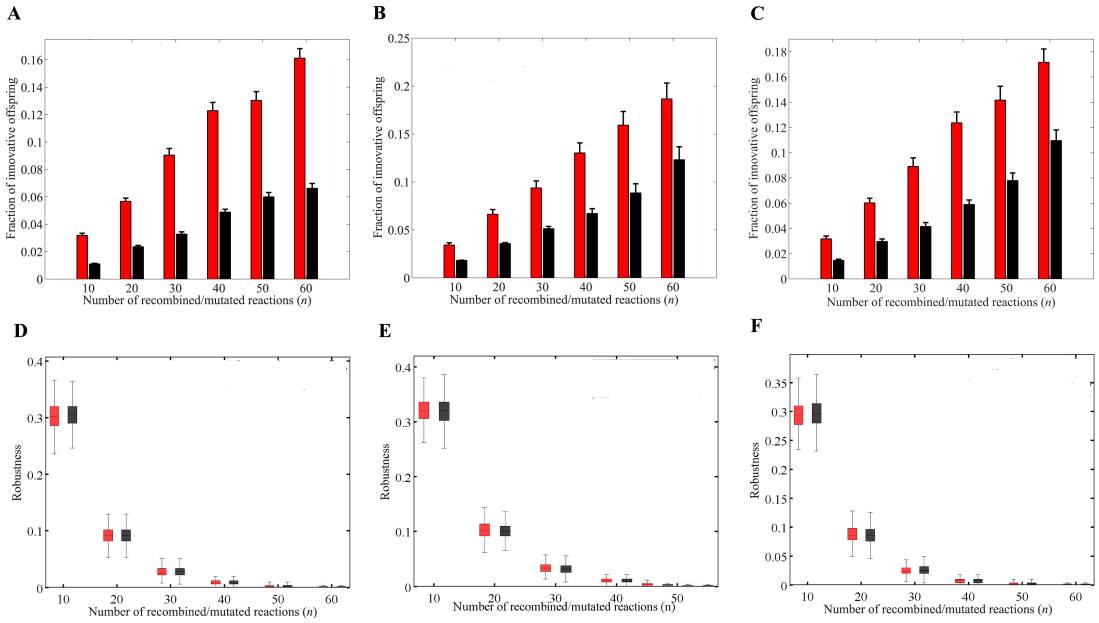
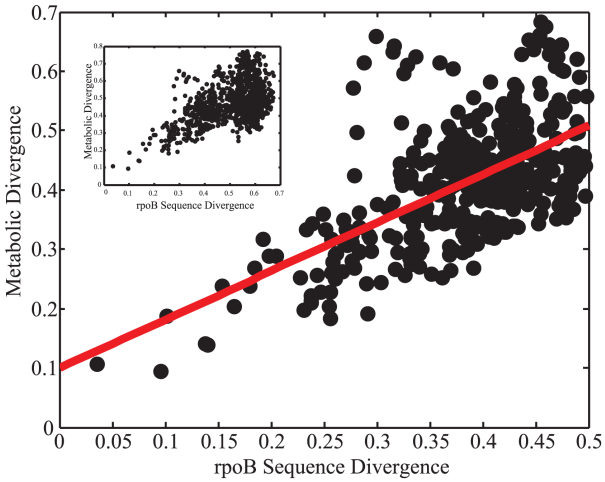
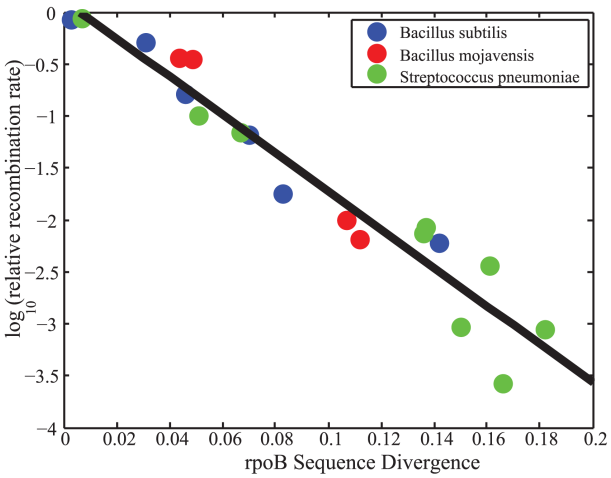


Figure S3: Vertical axes in panels **(a), (b),** and **(c)** show mean (bars) and standard error (vertical lines) of the fraction of innovative offspring (f_{innov}) generated by recombination (red) versus mutation (black), as a function of the number of reaction changes (n, x-axis) among those offspring retaining viability respectively on (a) glucose, (b) acetate, and (c) acetate. Parental metabolic network pairs are sampled based on the (a) second, (b) second, and (c) first approach (See texts S1e, and S2). Vertical axes in panels **(d), (e),** and **(f)** show recombinational robustness (red) versus mutational robustness (black) , that are defined as the fraction of recombinant (or mutant) offspring retaining viability respectively on (d) glucose, (e) acetate, and (f) acetate. Parental metabolic network pairs are sampled based on the (d), second (e) second, and (f) first approach (See texts S1f, and S2). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

A


Metabolic Divergence
rpoB Sequence Divergence
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7

B


Bacillus subtilis
Bacillus mojavensis
Streptococcus pneumoniae
log10 (relative recombination rate)
rpoB Sequence Divergence
0
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2

C

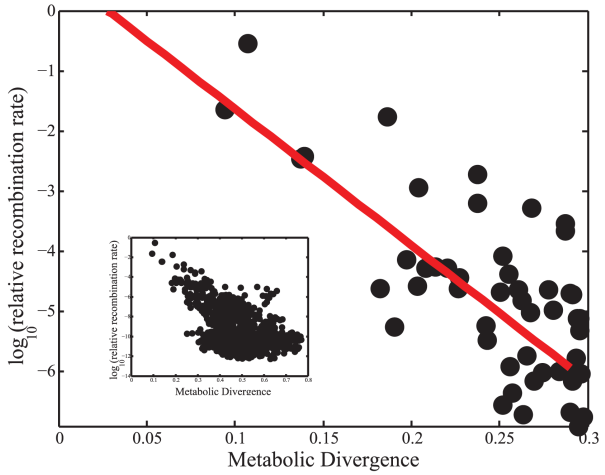

log10 (relative recombination rate)
Metabolic Divergence
0
−1
−2
−3
−4
−5
−6
0.05
0.1
0.15
0.2
0.25
0.3

Figure S4: Relative recombination rates. (a) Metabolic divergence, defined as the normalized Hamming distance between the genotype vectors of two metabolic networks, is correlated (Pearson's $r=0.60$, $P<10^{-40}$) with sequence divergence, defined as the normalized Hamming distance between rpoB (RNA polymerase) sequences of the corresponding pair of species. Each point corresponds to one of $\binom{51}{2}$ possible species pairs chosen from 51 distinct species (inset) whose pairwise rpoB sequence divergence lies below 0.5 [49]. **(b)** Relative rate of recombination, for a range of related donor species as a function of sequence divergence for a variety of bacterial recipients: *Bacillus subtilis* (blue), *Bacillus mojavensis* (red), and *Streptococcus pneumoniae* (green). The best log-linear fit is shown (black line), with an intercept of 0.11 and a slope of -18.40. Data is based on [50,51]. **(c)** Relative recombination rate (logarithmic scale, *y*-axis) as a function of metabolic divergence (*x*-axis) for metabolic network pairs with metabolic divergence lower than 0.3, chosen among the set of all possible metabolic network pairs (inset). The red line is the result of a linear regression with a regression coefficient of -22.57, and an intercept of 0.62. Data in (c) is based on the linear relationship from (b), and the metabolic divergence of (a).

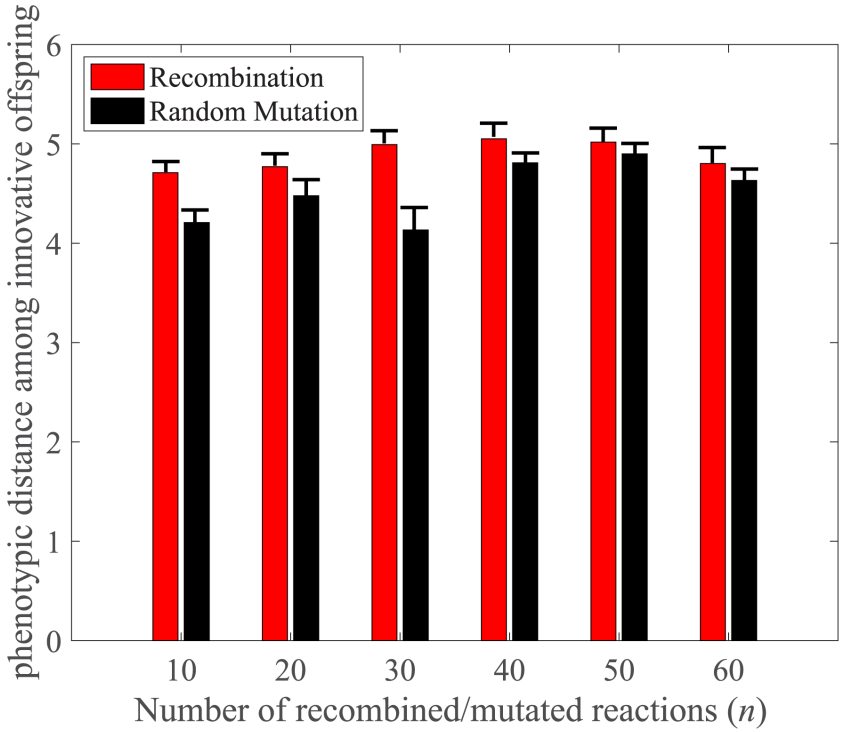


Figure S5: Phenotypic diversity among innovative offspring. Vertical axis shows mean (bars) and standard error (vertical lines) of the phenotypic distance among all pairs of innovative offspring generated by recombination (red) versus mutation (black), as a function of the number of reaction changes (n, x-axis). Phenotypic distance (ΔP) between a given pair of innovative offspring is measured as the number of carbon sources on which only one offspring but not the other is viable on.

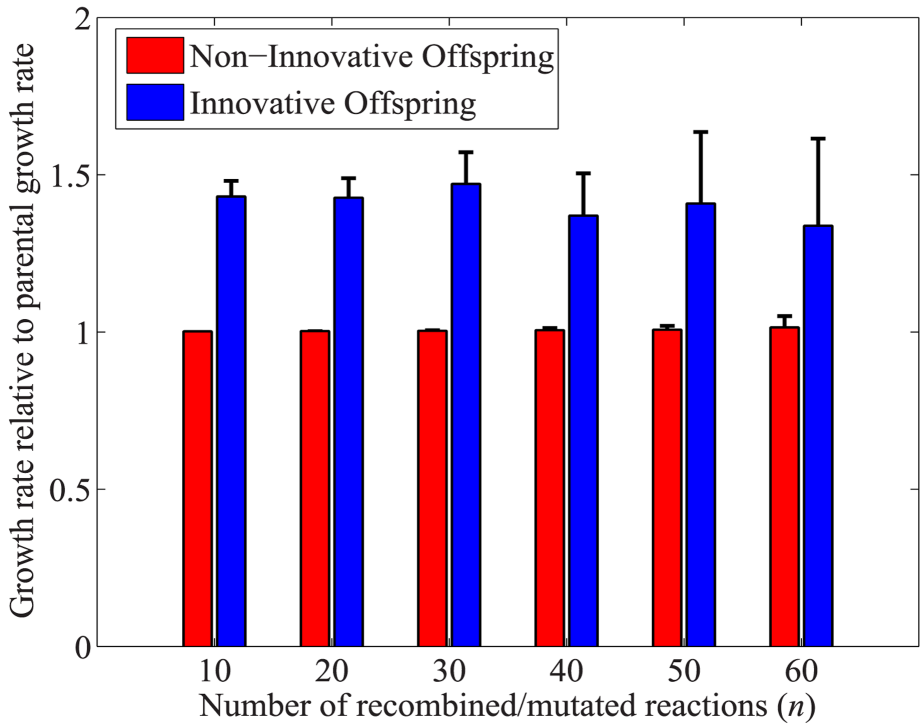
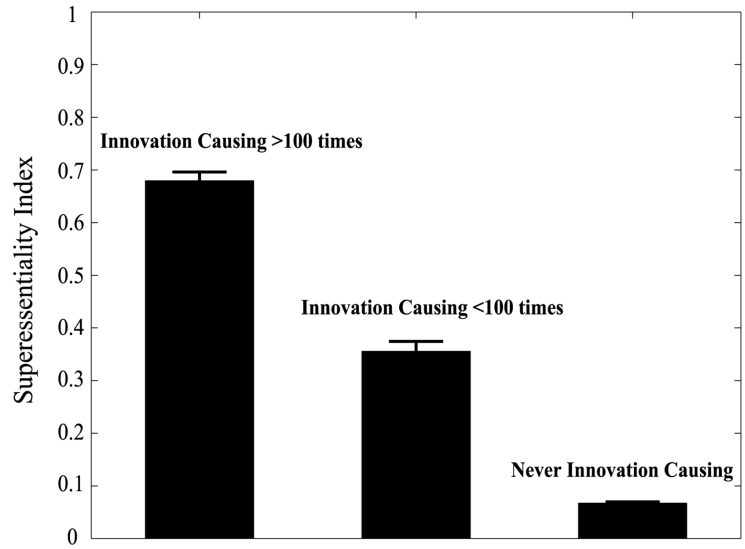
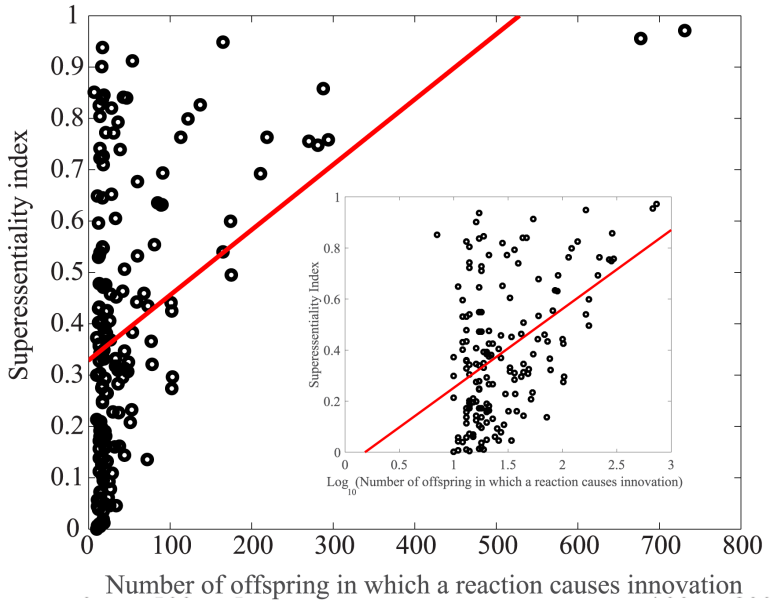


Figure S6: Mean (bar) and standard error (vertical line) of the biomass growth flux of innovative offspring (blue), and non-innovative offspring (red), divided by the parental growth rate, as a function of the number of recombined reactions (*n*). For this analysis, we created 1000 random metabolic network pairs viable only on glucose and with a difference in growth rate less than 0.25 percent. Regardless of *n*, the relative growth rate of non-innovative offspring is approximately equal to one, meaning that their growth rate is equal to the parental growth rate. In contrast, the relative growth rate for innovative offspring exceeds 1.4 for all *n*, and so innovative offspring even on the original carbon source can grow more than 40 percent faster than their parents.

A


Superessentiality Index
Innovation Causing >100 times
Innovation Causing <100 times
Never Innovation Causing

B


Superessentiality index
Number of offspring in which a reaction causes innovation
Superessentiality Index
Log10(Number of offspring in which a reaction causes innovation)

C

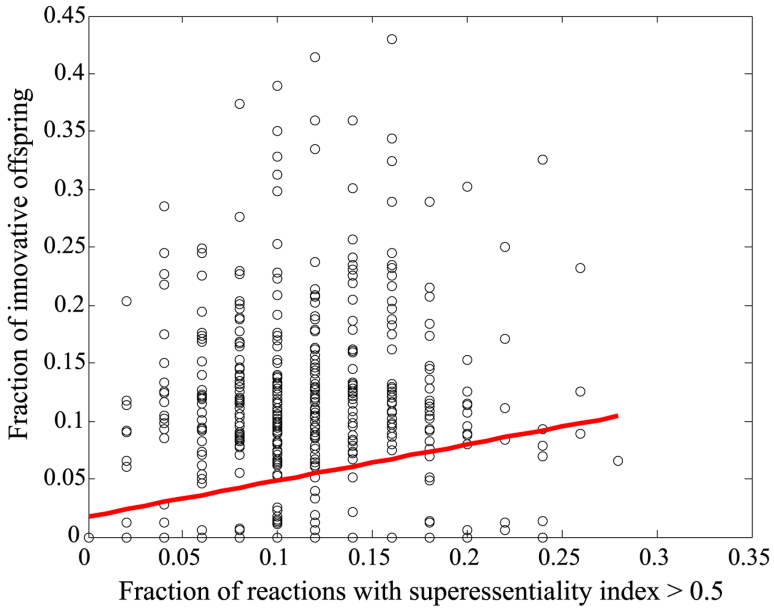

Fraction of innovative offspring
Fraction of reactions with superessentiality index > 0.5

Figure S7: Superessential reactions and metabolic innovation. (a) Mean (bars) and standard error (vertical lines) of the superessentiality index of reactions that (i) cause innovation in more than 100 innovative offspring (left), (ii) cause innovation in fewer than 100 innovative offspring (middle), and (iii) never cause innovation (right). **(b)** Scatterplot of superessentiality index (I_{SE}; y-axis) versus number of innovations (x-axis) caused by innovation-causing reactions (positive correlation: (Pearson's $r=0.47$, $P<10^{-9}$)). Horizontal axis in the inset is shown in logarithmic scale to improve visual clarity. **(c)** The fraction of innovative recombinant offspring (f_{innov}, y-axis) is significantly correlated (Pearson's $r=0.18$, $P<10^{-8}$) with the fraction of reactions with superessentiality index higher than 0.5 (f_{super}, x-axis) among reactions that can potentially be transferred from donor to the recipient.

When studying metabolic innovation, it is important to distinguish two classes of reactions with high superessentiality index. The first comprises reactions with superessentiality index I_{SE}=1, which are needed in all viable metabolic networks [59]. These reactions are crucial for retaining viability on parental carbon sources, but they play no role in metabolic innovation, because all metabolic networks must have them. The second class includes reactions where $0.5< I_{SE} <1$. These reactions are less crucial for retaining viability on parental carbon sources, but important for gaining viability on novel carbon sources. They can be absent in some metabolic networks, because metabolic pathways that by-pass them exist, which means that they can be involved in recombinational exchange, and thus in the origin of novel phenotypes. Our analysis above highlights the special importance of these reactions for metabolic innovation.

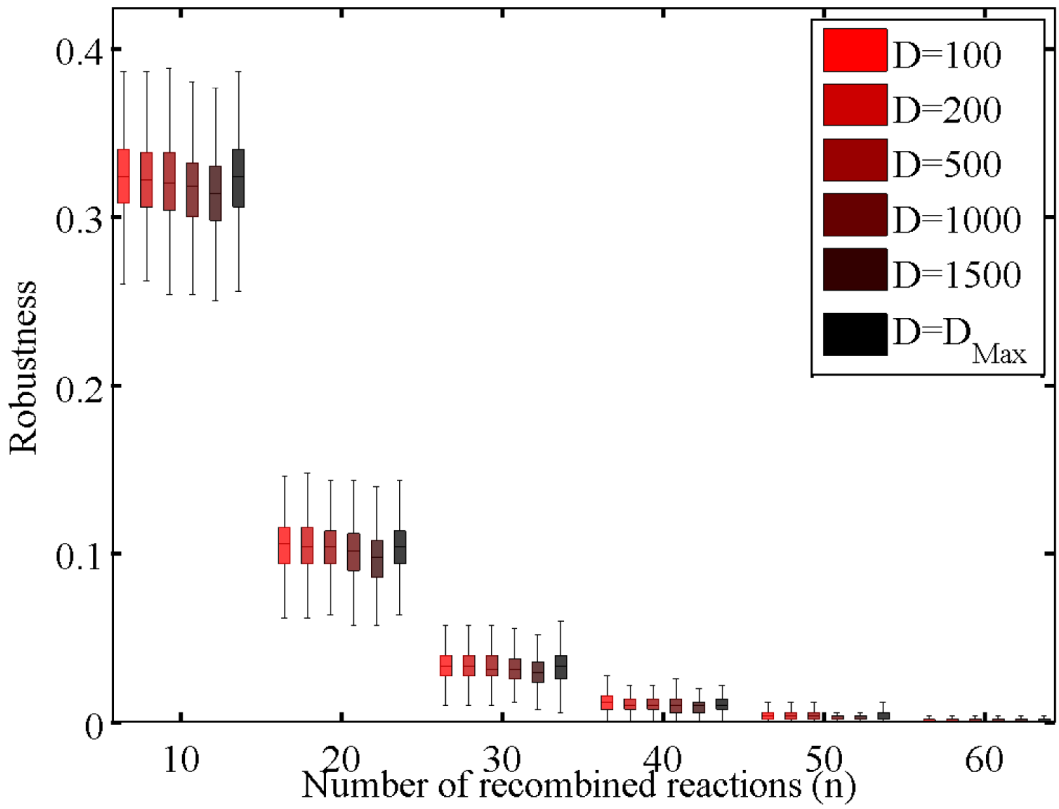


Figure S8: Effect of parental genotypic diversity on recombinational robustness. The vertical axis shows recombinational robustness, that is, the fraction of offspring that retain viability on glucose and that are generated by recombination between parental metabolic networks with genotypic distance (D), where D is color-coded according to the legend. The horizontal axis shows the number of recombined reactions (n). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

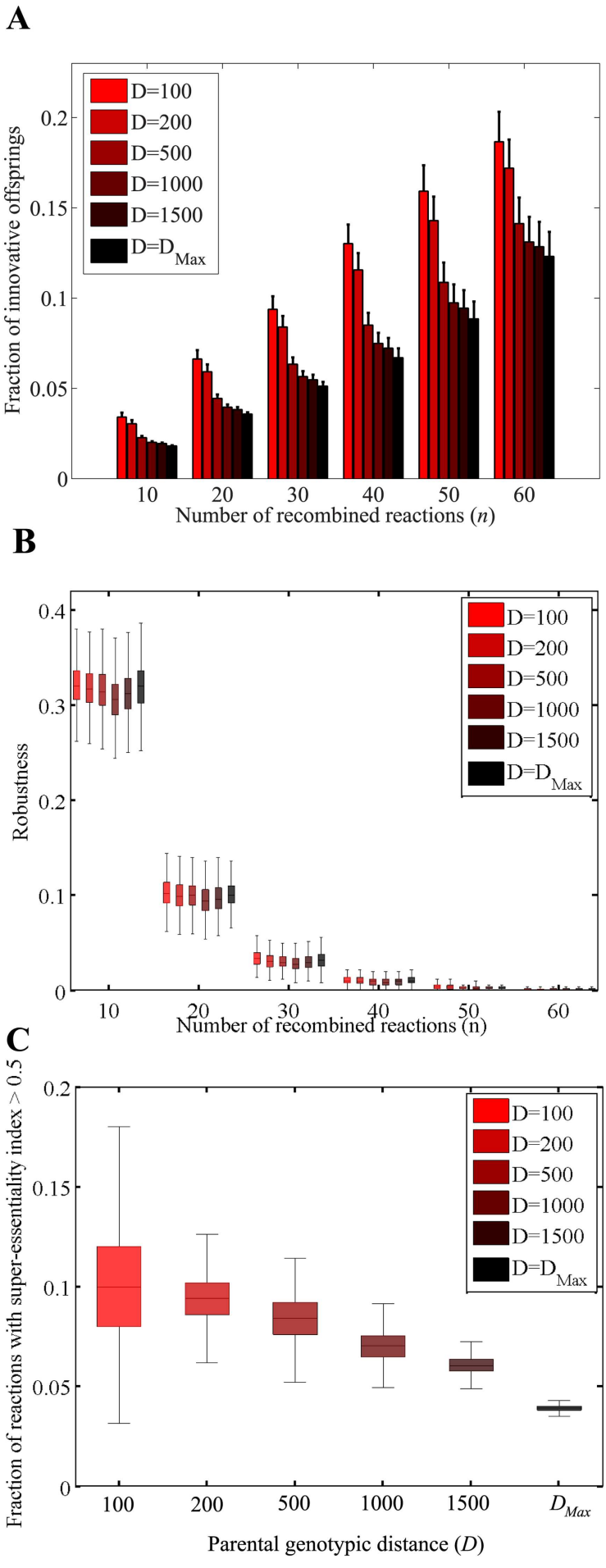

A
D=100
D=200
D=500
D=1000
D=1500
D=D Max
Fraction of innovative offsprings
0
0.05
0.1
0.15
0.2
10
20
30
40
50
60
Number of recombined reactions (n)
B
D=100
D=200
D=500
D=1000
D=1500
D=D Max
Robustness
0
0.1
0.2
0.3
0.4
10
20
30
40
50
60
Number of recombined reactions (n)
C
D=100
D=200
D=500
D=1000
D=1500
D=D Max
Fraction of reactions with super-essentiality index > 0.5
0
0.05
0.1
0.15
0.2
100
200
500
1000
1500
D Max
Parental genotypic distance (D)

Figure S9: Effect of parental genotypic diversity on recombinational innovation and robustness (parental metabolic networks are required to be viable on acetate). (a) the mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}), generated by recombination between parental metabolic networks viable on acetate with genotypic distance (D), where D is color-coded according to the legend. The horizontal axis shows the number of recombined reactions (n). **(b)** The vertical axis shows recombinational robustness, that is, the fraction of offspring that retain viability on acetate and are generated by recombination between parental metabolic networks with genotypic distance (D), where D is color-coded according to the legend of panel (a). The horizontal axis shows the number of recombined reactions (n). **(c)** The fraction of reactions with superessentiality index higher than 0.5 (f_{super}, x-axis) among reactions that can potentially be transferred from the parental donor to the recipient metabolic network, with genotypic distance (D, x-axis). Note that parental metabolic networks are required to be viable on acetate instead of glucose. All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

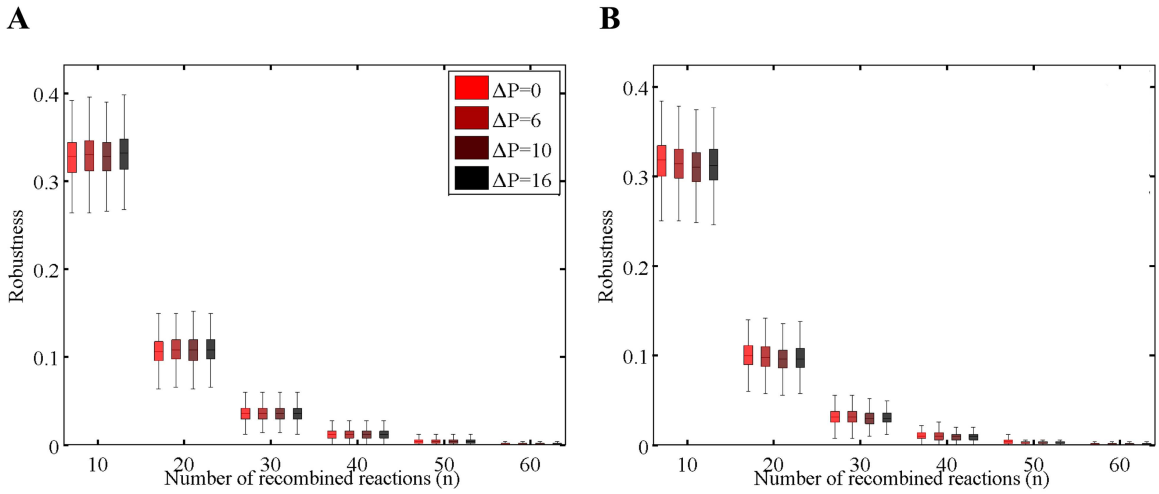


Figure S10: Effect of parental phenotypic diversity on recombinational robustness. The vertical axes show **(a)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in **(b)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity ΔP (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions (n). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

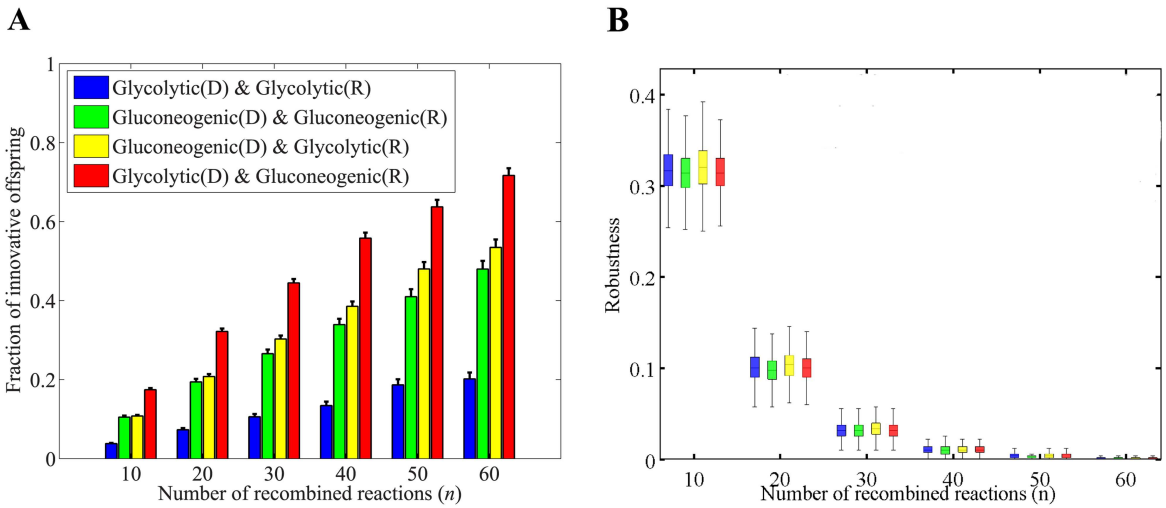


Figure S11: Effect of parental carbon source classes on metabolic innovation. The vertical axes in **(a)** show the mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}), and in **(b)** the fraction of recombinant offspring retaining viability on glucose (for those recipients viable on glycolytic carbon sources) or acetate (for those recipients viable on gluconeogenic carbon sources). The horizontal axes show the number of recombined reactions (n). For this analysis we generated offspring by recombination between parental metabolic networks in which (i) the donor was viable on 5 glycolytic carbon sources and the recipient was viable on 5 other glycolytic carbon sources (blue), (ii) the donor was viable on 5 gluconeogenic carbon sources and the recipient was viable on 5 other gluconeogenic carbon sources (green), (iii) the donor was viable on 5 gluconeogenic carbon sources and the recipient was viable on 5 glycolytic carbon sources (yellow), and (iv) the donor was viable on 5 glycolytic carbon sources and the recipient was viable on 5 gluconeogenic carbon sources (red). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

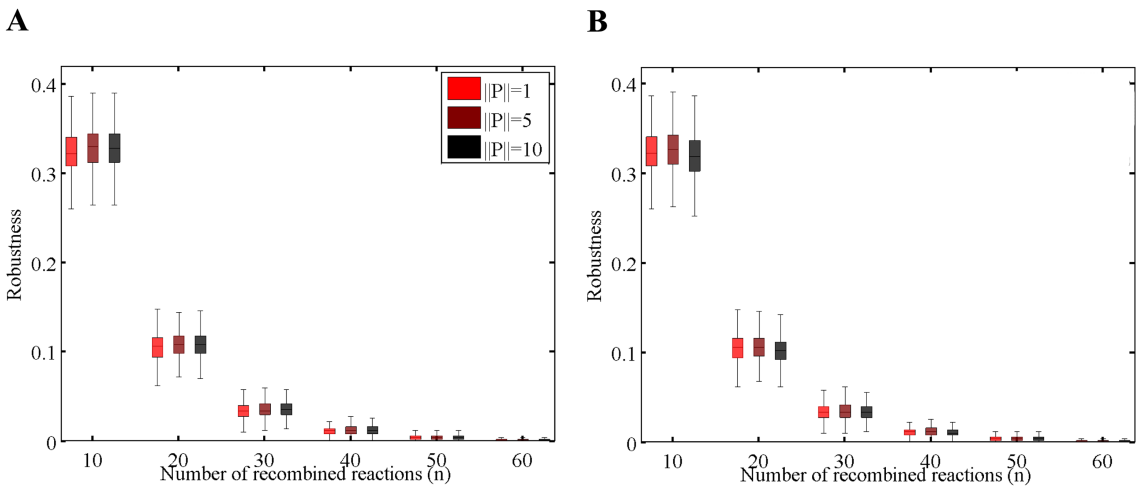


Figure S12: Effect of parental phenotypic complexity on metabolic robustness (for $||P|| \leq 10$, and $\Delta P=0$). The vertical axes show **(a)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in (**b)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity $||P||$ (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions (n). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima

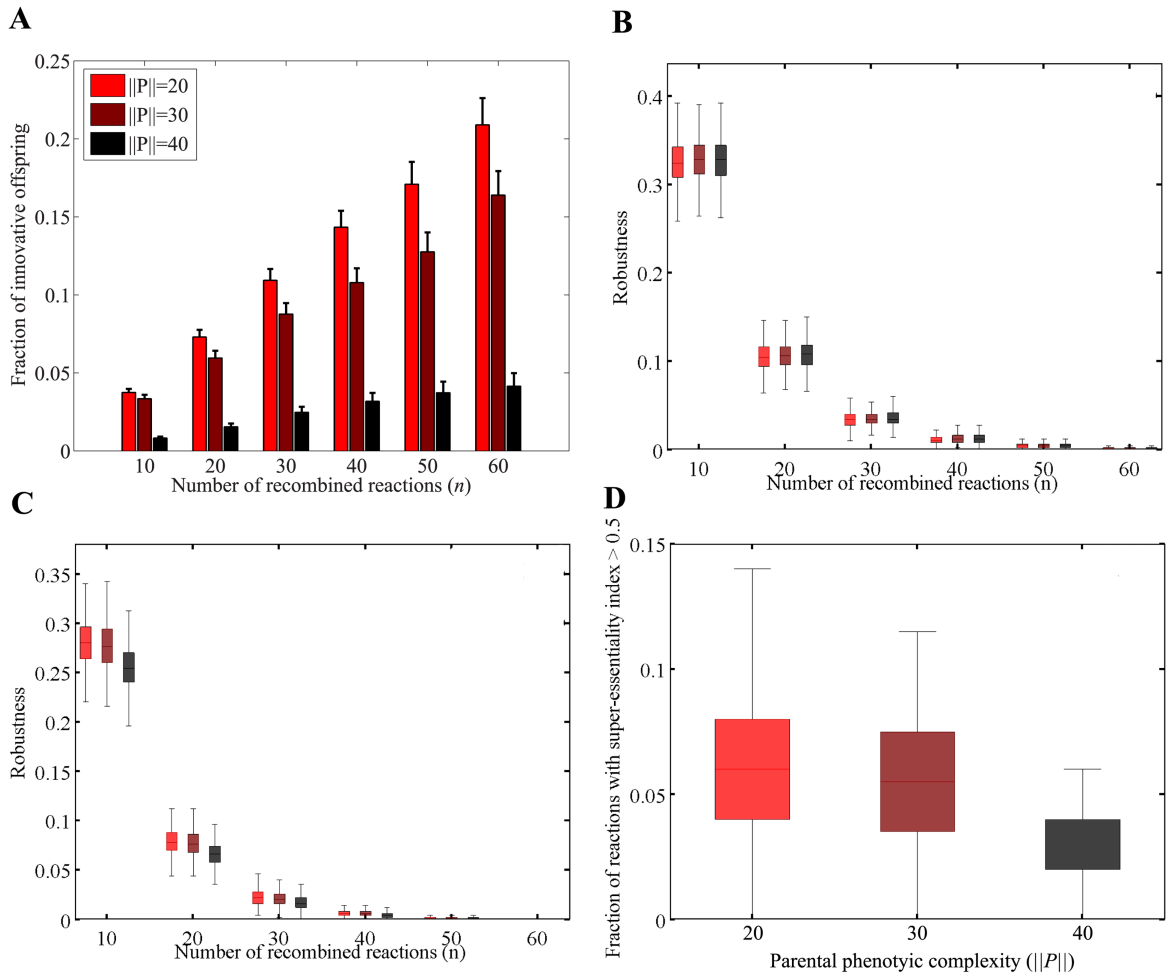


Figure S13: Effect of parental phenotypic complexity on metabolic innovation (for $||P|| > 10$, and *ΔP=10*). The vertical axes show in **(a)** the mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}), in **(b)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in **(c)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity $||P||$ (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions (n). **(d)** Distribution of the fraction of reactions with superessentiality index exceeding 0.5 (y-axis) among the reactions that can potentially be transferred from the parental donor metabolic network to the recipient, with phenotypic complexity ($||P||$, x-ais). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

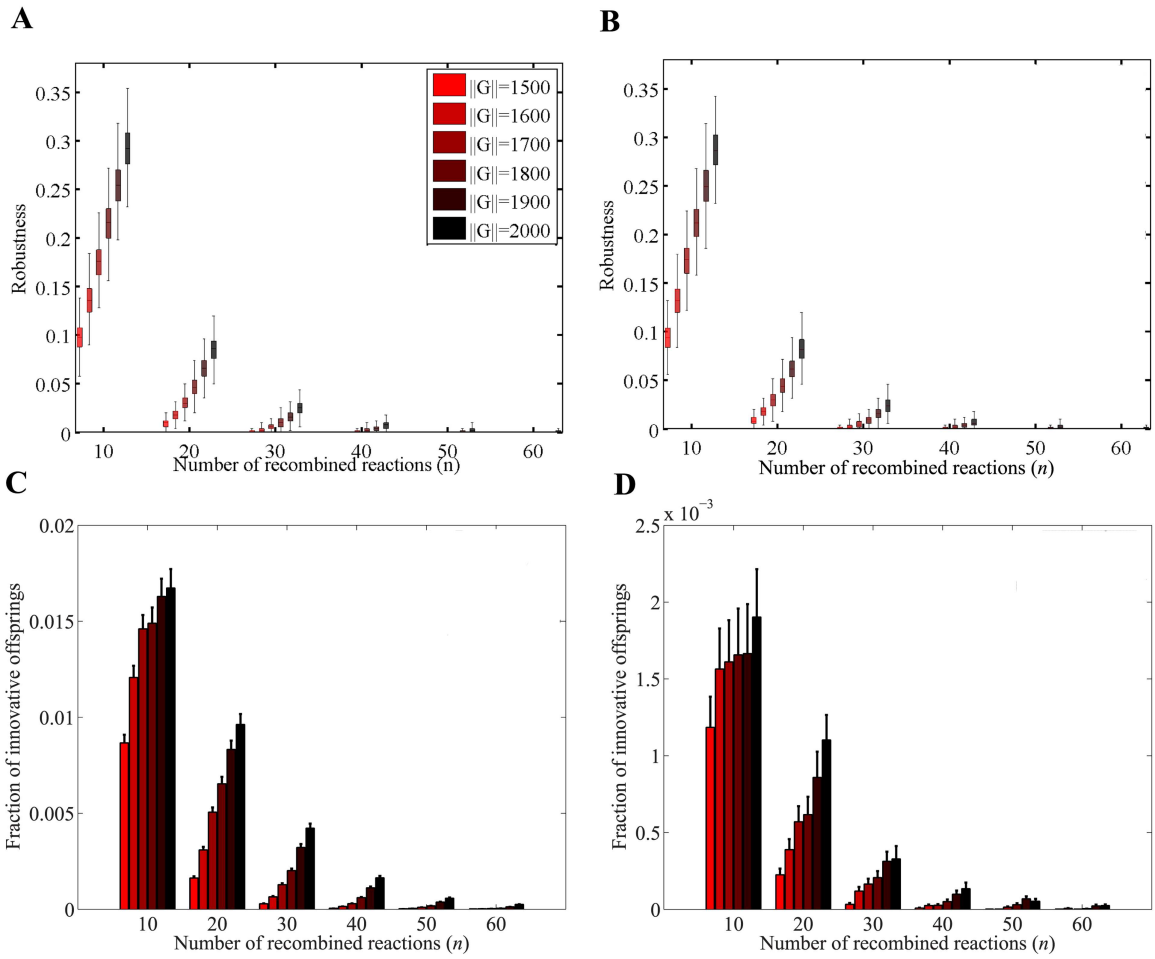


Figure S14: Effect of genotypic complexity (metabolic network size *(||G||)*) on recombinational innovation. Vertical axes in panels **(a)**, and **(b)** show the fraction of recombinant offspring retaining viability (i.e. robustness, vertical axis), on (a) glucose, and (b) acetate, are shown as a function of the number of recombined reactions (*n*). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima. Panels **(c)**, and **(d)** show mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}), generated by recombination between parental metabolic networks required to be viable on (c) glucose, and (d) acetate, with size *(||G||)* color-coded as in the legend, are shown as a function of the number of recombined reactions (*n*).

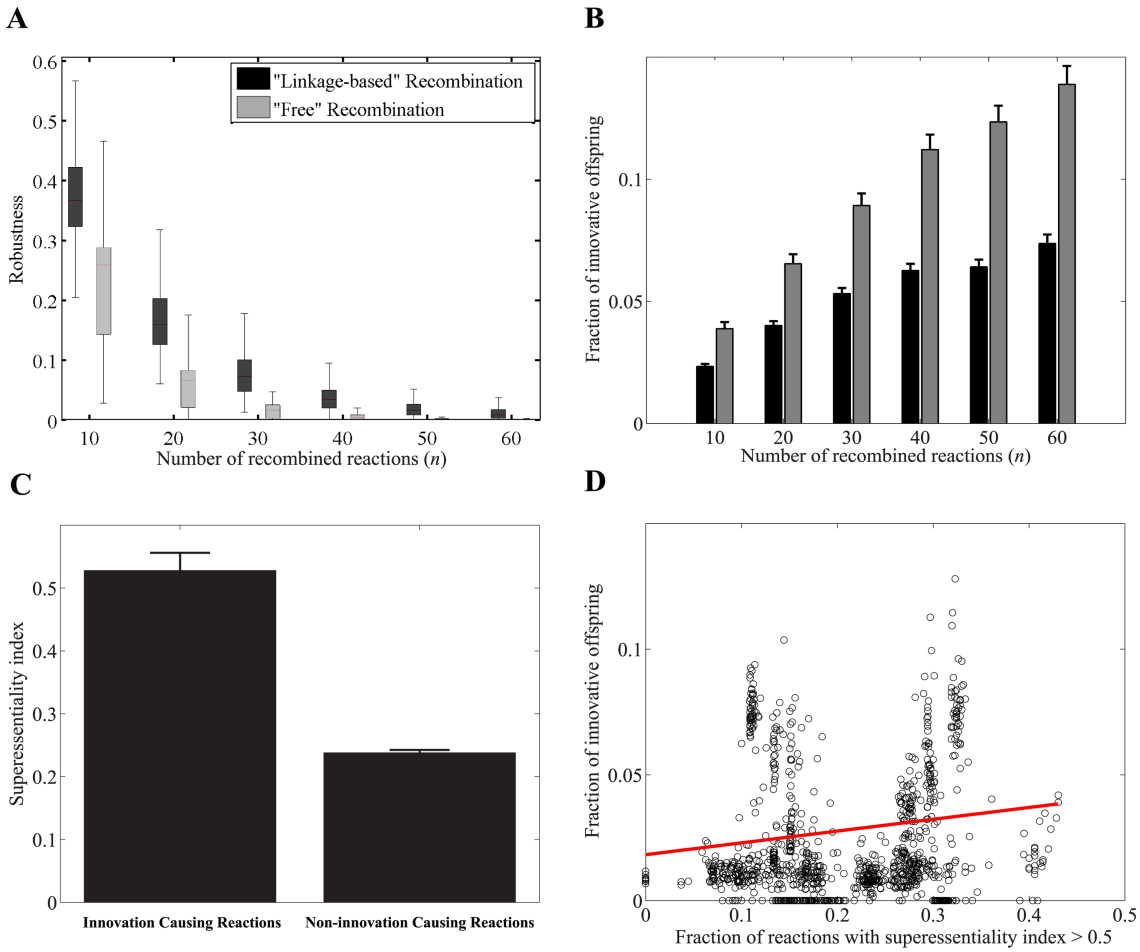


Figure S15: (a) Fraction of robust recombinant offspring, i.e., offspring retaining viability on all the carbon sources that the recipient parental metabolic network is viable on (*y*-axis), as a function of the number of recombined reactions (x-axis). Offspring were generated by i) linkage-based recombination between prokaryotic metabolic networks (black), and ii) free recombination between prokaryotic metabolic networks (gray). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima**. (b)** Mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}) generated by (i) linkage-based recombination (black), and (ii) free recombination between prokaryotic metabolic networks (gray)**. (c)** Mean (bars) and standard error (vertical lines) of the superessentiality index of reactions that cause innovation (left) as compared with those never causing innovation (right). **(d)** The fraction of innovative recombinant offspring (f_{innov}, *y*-axis) is significantly associated (Pearson's $r=0.13$, $P<10^{-5}$) with the fraction of reactions with superessentiality index higher than 0.5 (*x*-axis) among the set of reactions that can potentially be transferred from the parental donor to the recipient metabolic network.

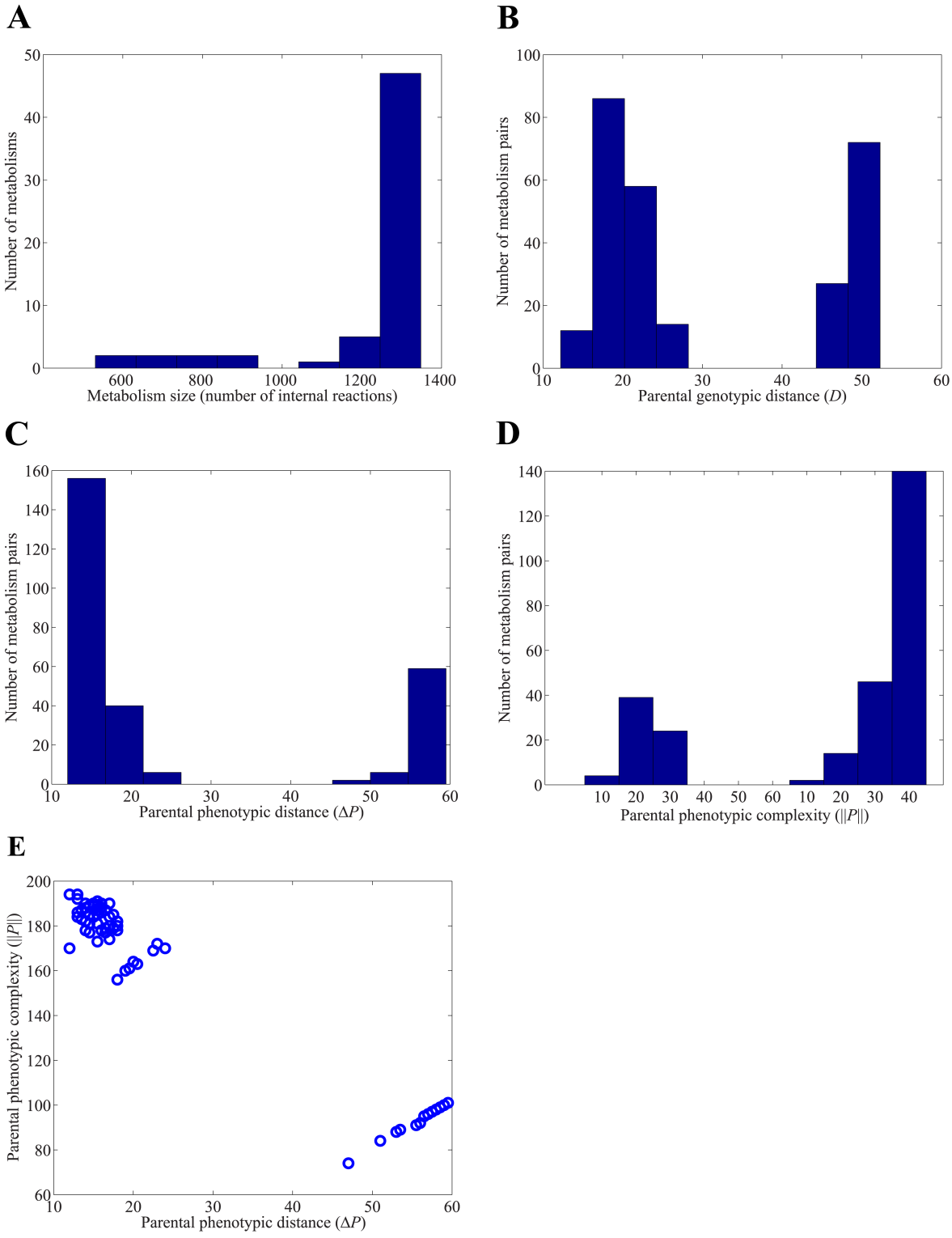
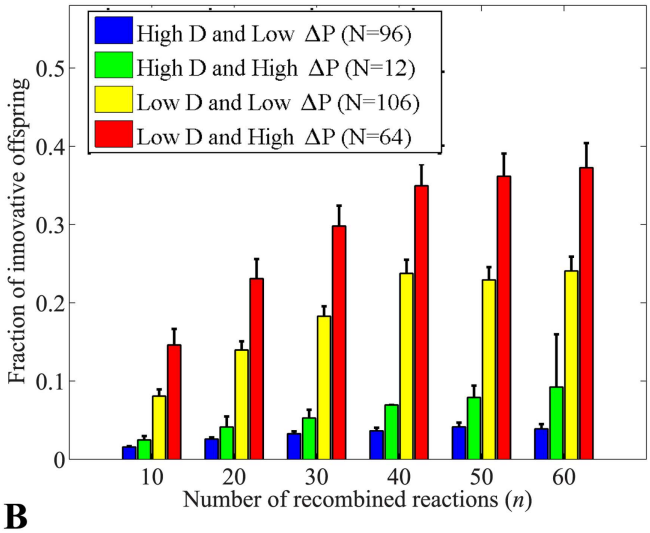
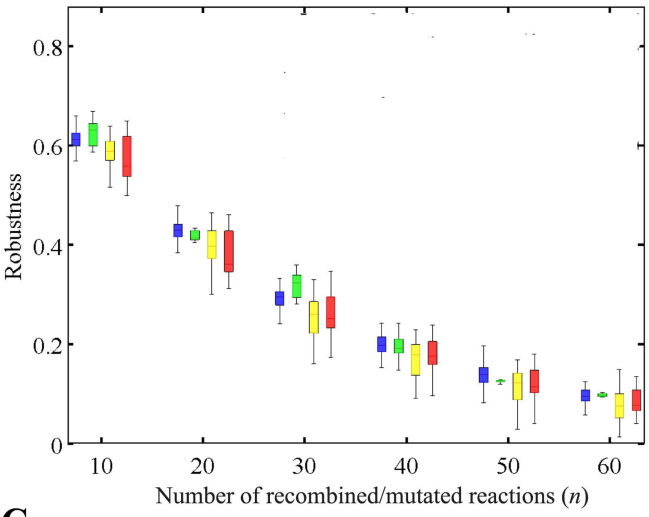


Figure S16: Distribution of parental genotypic and phenotypic features among prokaryotic metabolic networks. (a) Histogram of the number of metabolic networks with a given metabolic network size (approximated by the number of internal reactions) specified on the *x*-axis. Vertical axes in panels **(b)**, **(c)**, and **(d)** show the number of parental metabolic network pairs with a given b) genotypic distance (D), c) phenotypic distance (ΔP), and d) phenotypic complexity ($||P||$), as specified on the *x*-axes. **(e)** Each circle represents a given parental metabolic network pairs with a given phenotypic distance (ΔP, *x*-axis), and phenotypic complexity ($||P||$, *y*-axis).

A


High D and Low ΔP (N=96)
High D and High ΔP (N=12)
Low D and Low ΔP (N=106)
Low D and High ΔP (N=64)
Fraction of innovative offspring
0
0.1
0.2
0.3
0.4
0.5
10
20
30
40
50
60
Number of recombined reactions (n)

B


Robustness
0
0.2
0.4
0.6
0.8
10
20
30
40
50
60
Number of recombined/mutated reactions (n)

C

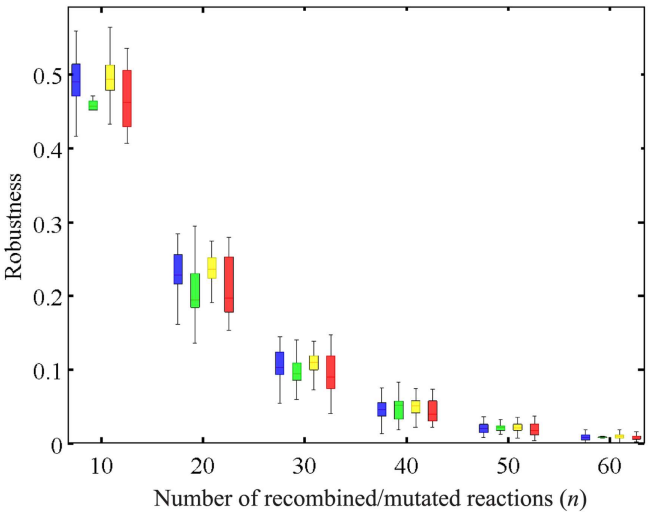

Robustness
0
0.1
0.2
0.3
0.4
0.5
10
20
30
40
50
60
Number of recombined/mutated reactions (n)

Figure S17: The vertical axes show **(a)** mean (bar) and standard error (vertical line) of the fraction of innovative offspring (f_{innov}), **(b)** robustness to linkage-based, and **(c)** robustness to "free"recombination. Here we define robustness as the fraction of recombinant offspring retaining viability on at least one of the carbon source(s) on which the parental recipient metabolic networks are viable. Offspring are generated by recombination between prokaryotic parental metabolic networks with i) high genotypic distance ($D>40$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (blue, $N=96$ parental pairs), ii) high genotypic distance ($D>40$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (green, $N=12$ parental pairs), iii) low genotypic distance ($D<30$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (yellow, $N=106$ parental pairs), and iv) low genotypic distance ($D<30$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (red, $N=64$ parental pairs).

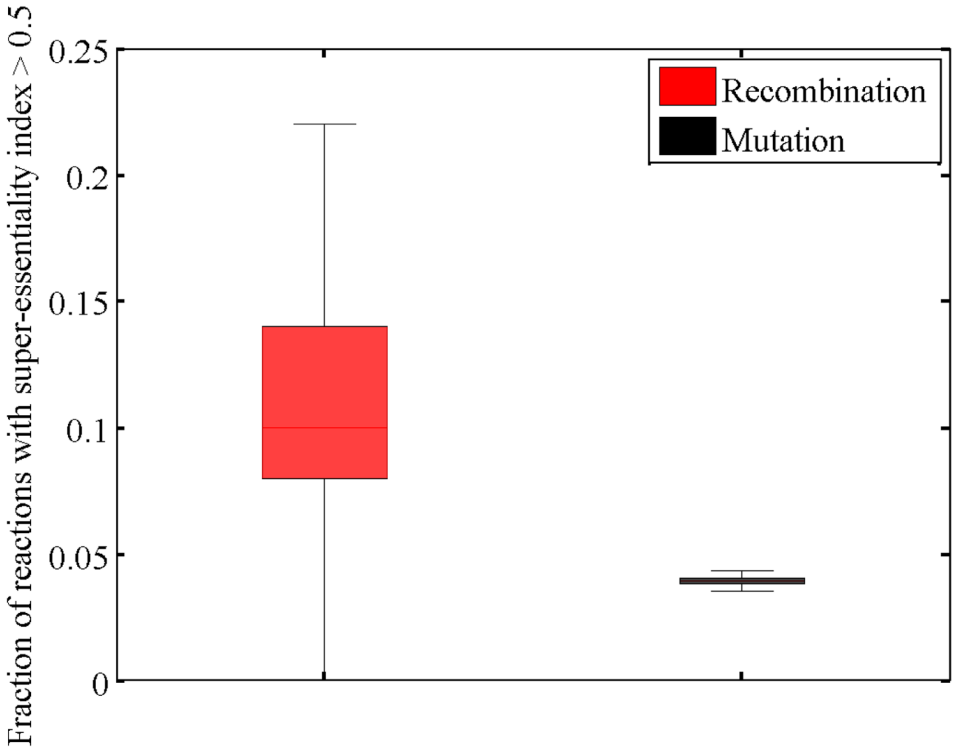


Figure S18: The fraction of reactions with superessentiality index higher than 0.5 (f_{super}, x-axis) among the set of reactions that can potentially be transferred to the recipient metabolic network via recombination (red box) versus random mutation (black box). Boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.